







EDUCATIONAL and  
PSYCHOLOGICAL  
MEASUREMENT

A QUARTERLY JOURNAL DEVOTED TO THE DEVELOPMENT AND APPLICATION OF MEASURES OF INDIVIDUAL DIFFERENCES

<i>Preschool Racial Attitude Measure II.</i> JOHN E. WILLIAMS, DEBORAH L. BEST, DONNA A. BOSWELL, LINDA A. MATTSO, AND DEBORAH J. GRAVES .....	3
<i>Strategic Use of Random Subsample Replication and a Coefficient of Factor Replicability.</i> WILLIAM G. KATZENMEYER AND A. JACKSON STENNER .....	19
<i>Domain Validity and Generalizability.</i> HENRY F. KAISER AND WILLIAM B. MICHAEL .....	31
<i>An Empirical Investigation Comparing the Effectiveness of Four Scoring Strategies on the Kuder Occupational Interest Survey Form DD.</i> STEPHEN OLEJNIK AND ANDREW C. PORTER .....	37
<i>The 27 Percent Rule Revisited.</i> RALPH B. D'AGOSTINO AND EDWARD E. CURETON .....	47
<i>A Model for Psychometrically Distinguishing Aptitude from Ability.</i> SUSAN E. WHITELY AND RENÉ V. DAWIS .....	51
<i>A Measure of the Average Intercorrelation.</i> EDWARD P. MEYER .....	67
<i>Effects of a Confidence Weighted Scoring System on Measures of Test Reliability and Validity.</i> RICHARD C. PUGH AND J. JAY BRUNZA ..	73
<i>A Test for Homogeneity of Regression without Homogeneity of Variance.</i> DONALD H. McLAUGHLIN .....	79
<i>Convergent and Divergent Measurement of Creativity in Children.</i> WILLIAM C. WARD .....	87
<i>Longitudinal Studies of Risk Taking on Objective Examinations.</i> MALCOLM J. SLAKTER, KEVIN D. CREHAN, AND ROGER A. KOEHLER ..	97
<i>The Effect of Double Standardized Scoring on the Semantic Differential.</i> JACK R. HAYNES .....	107
<i>Item-Analysis of Jourard's Self-Disclosure Questionnaire-21.</i> W. BARNETT PEARCE AND BERNIE WIERE .....	115

(Continued on inside front cover)



(Continued from front cover)

<i>The Classroom Boundary Questionnaire: An Instrument to Measure One Aspect of Teacher Leadership in the Classroom.</i> THOMAS L. MORRISON .....	119
<i>An Assessment of the Effectiveness of Complex Alternatives in Multiple Choice Achievement Test Items.</i> DANIEL J. MUELLER .....	135

## COMPUTER PROGRAMS

<i>The Analysis of Multivariate Group Differences.</i> ALAN L. GROSS ....	143
<i>Computer Programs for Robust Analyses in Multifactor Analysis of Variance Designs.</i> PAUL A. GAMES .....	147
<i>A FORTRAN Program for Simulating Educational Growth with Varying School Impact.</i> JAMES M. RICHARDS, JR., NANCY KARWEIT, AND TRUMAN W. PREVATT .....	153
<i>A Computer Program to Test a Repeated Measures Hypothesis Using Hotelling's One-Sample <math>T^2</math> Statistic.</i> PETER P. VITALIANO AND SILAS HALPERIN .....	159
<i>LPA2: A FORTRAN V Computer Program for Green's Solution of Latent Class Analysis Applied to Latent Profile Analysis.</i> BERTIL MÅRDBERG .....	163
<i>A Population Subgroup Multiple Comparison Computer Program Based upon Categorical Data.</i> BERNARD A. RAFACZ .....	167

(Continued on outside back cover)

This journal is open to: (1) discussions of problems in the field of the measurement of individual differences, (2) reports of research on the development and use of tests and measurements in education, industry, and government, (3) descriptions of testing programs being used for various purposes, and (4) miscellaneous notes pertinent to the measurement field, such as suggestions of new types of items or improved methods of treating test data. Authors are granted permission to have reprints made of their own articles for their own use at their own expense. Manuscripts should be sent in duplicate to Dr. W. Scott Gehman, Box 6907 College Station, Durham, North Carolina 27708. Authors are requested to put tables, footnotes, and abstracts on pages separate from the text and to follow the general directions given in the *Publication Manual of the American Psychological Association (1974 Revision)*. Journal titles should not be abbreviated.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT is published quarterly, one volume per calendar year, at 3121 Cheek Road, Durham, North Carolina 27704. Second class postage paid at Durham, North Carolina and other cities.

Publication charges to authors are as follows: \$30.00 per page of running text; \$40.00 per page of tables, figures, and formulas.

Subscription rate, \$16.00 a year, domestic and foreign. Single copies, \$4.00. Back volumes: Volumes 5 to the present \$16.00 each. See inside back cover for information relative to back issues.

Orders should be sent to EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, Box 6907, College Station, Durham, North Carolina 27708.

Copyright © 1975 by Frederic Kuder



# EDUCATIONAL and PSYCHOLOGICAL MEASUREMENT

QUARTERLY JOURNAL DEVOTED TO THE DEVELOPMENT AND  
APPLICATION OF MEASURES OF INDIVIDUAL DIFFERENCES

28/8/77  
Booked &  
SSN to his  
P. R. ...

<i>Configural Frequency Analysis as a Statistical Tool for Defining Types.</i> G. A. LIENERT AND J. KRAUTH .....	231
<i>A Method for Hierarchical Clustering of a Matrix of a Thousand by a Thousand.</i> LOUIS L. MCQUITTY AND VALERIE L. KOCH .....	239
<i>Analysis Techniques for Exploratory Use of the Multitrait-Multimethod Matrix.</i> MICHAEL L. RAY AND ROGER M. HEELER .....	255
<i>Behavior of the Product-Moment Correlation Coefficient when Two Heterogeneous Subgroups Are Pooled.</i> ALAN L. SOCKLOFF .....	267
<i>The r-Point Biserial Limitation.</i> R. A. KARABINUS .....	277
<i>On Splitting the Tails Unequally: A New Perspective on One-versus Two- Tailed Tests.</i> SANFORD L. BRAVER .....	283
<i>A Non-Parametric Test for Increasing Trend.</i> MARTIN COOPER .....	303
<i>The Variances of Empirically Derived Option Scoring Weights.</i> GARY ECHTERNACHT .....	307
<i>The Application of Discriminant Function Analysis to Correlated Sam- ples.</i> G. FRANK LAWLIS AND ARTHUR B. SWENEY .....	313
<i>A Comparison of Variable Configurations across Scale Lengths: An Em- pirical Study.</i> HOWARD G. SCHUTZ AND MARGARET H. RUCKER .....	319
<i>An Investigation of the Rasch Simple Logistic Model: Sample Free Item and Test Calibration.</i> HOWARD E. A. TINSLEY AND RENÉ V. DAWIS .....	325
<i>Attempt to Construct a Scale for the Measurement of the Effect of Sug- gestion on Perception.</i> V. A. GHEORGHIU, V. HODAPP, AND C. M. LUDWIG .....	341
<i>A Study of the Effect of the Violation of the Assumption of Independent Sampling upon the Type I Error Rate of the Two-Group t-Test.</i> ROBERT W. LISSITZ AND STEVE CHARDOS .....	353

## VALIDITY STUDIES OF ACADEMIC ACHIEVEMENT

<i>The Relative Validity of Scales Prepared by Naive Item Writers and Those Based on Empirical Methods of Personality Scale Construction.</i> DOUGLAS N. JACKSON .....	361
---	-----

(Continued on inside front cover)

VOLUME THIRTY-FIVE, NUMBER TWO, SUMMER 1975



<i>Improving the Validity of Affective Self-Report Measures through Constructing Personality Scales Unconfounded with Social Desirability: A Study of the Personality Research Form.</i> ROBERT D. ABBOTT .....	371
<i>The Reliability and Validity of Two Objective Measures of Achievement Motivation for Adolescent Females.</i> MICHAEL POMERANTZ AND CHARLES B. SCHULTZ .....	379
<i>Prediction of College Achievement Using the Need Achievement Scale from the Edwards Personal Preference Schedule.</i> RONALD R. MORGAN .....	387
<i>Validity of the MMPI-168 for Psychiatric Screening.</i> JOHN E. OVERALL, JAMES N. BUTCHER, AND SARA HUNTER .....	393
<i>Comparison of the Standard MMPI and the Mini-Mult in a University Counseling Center.</i> R. B. SIMONO .....	401
<i>The Factorial Validity of the Piers-Harris Children's Self-Concept Scale for Each of Three Samples of Elementary, Junior High, and Senior High School Students in a Large Metropolitan School District.</i> WILLIAM B. MICHAEL, ROBERT A. SMITH, AND JOAN J. MICHAEL ....	405
<i>The Content and Construct Validity of the Barth Scale: Assumptions of Open Education.</i> ANTHONY J. COLETTA AND ROBERT K. GABLE .....	415
<i>Do These Co-Twins Really Live Together? An Assessment of the Validity of the Home Index as a Measure of Family Socio-economic Status.</i> LOUISE CARTER-SALTZMAN, SANDRA SCARR-SALAPATEK, AND WILLIAM B. BARKER .....	427
<i>Stability of Student Evaluations of Instructors and Their Courses with Implications for Validity.</i> HENRY J. OLES .....	437
<i>Protestant Ethic Attitudes among College Students.</i> L. K. WATERS, NICK BATLIS, AND CARRIE WHERRY WATERS .....	447
<i>Predictive Validity of the American University of Beirut Trial Aptitude Battery.</i> F. K. ABU-SAYF AND GEORGE I. ZA'ROUR .....	451

(Continued on outside back cover)

This journal is open to: (1) discussions of problems in the field of the measurement of individual differences, (2) reports of research on the development and use of tests and measurements in education, industry, and government, (3) descriptions of testing programs being used for various purposes, and (4) miscellaneous notes pertinent to the measurement field, such as suggestions of new types of items or improved methods of treating test data. Authors are granted permission to have reprints made of their own articles for their own use at their own expense. Manuscripts should be sent in duplicate to Dr. W. Scott Gehman, Box 6907 College Station, Durham, North Carolina 27708. Authors are requested to put tables, footnotes, and abstracts on pages separate from the text and to follow the general directions given in the *Publication Manual of the American Psychological Association (1974 Revision)*. Journal titles should not be abbreviated.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT is published quarterly, one volume per calendar year, at 3121 Cheek Road, Durham, North Carolina 27704. Second class postage paid at Durham, North Carolina and other cities.

Publication charges to authors are as follows: \$30.00 per page of running text; \$40.00 per page of tables, figures, and formulas.

Subscription rate, \$16.00 a year, domestic and foreign. Single copies, \$4.00. Back volumes: Volumes 5 to the present \$16.00 each. See inside back cover for information relative to back issues.

Orders should be sent to EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, Box 6907, College Station, Durham, North Carolina 27708.





# EDUCATIONAL and PSYCHOLOGICAL MEASUREMENT

A QUARTERLY JOURNAL DEVOTED TO THE DEVELOPMENT AND  
APPLICATION OF MEASURES OF INDIVIDUAL DIFFERENCES

Box  
to  
Docket 2  
Re  
Brain  
Das  
23/12/75

<i>Random Variables and Correlational Overkill.</i> JOSEPH T. KUNCE, DANIEL W. COOK, AND DOUGLAS E. MILLER .....	529
<i>Independence Problems for Certain Tests Based on the Shine-Bower Er- ror Term.</i> LESTER C. SHINE II .....	535
<i>On the Independence of Variable Sets.</i> CHARLES D. DZIUBAN, EDWIN C. SHIRKEY, AND THOMAS O. PEEPLES .....	539
<i>Sampling Characteristics of Kelley's <math>\epsilon^2</math> and Hays' <math>\omega^2</math></i> ROBERT M. CAR- ROLL AND LENA A. NORDHOLM .....	541
<i>Obtaining Paired Comparisons Data from Multiple Rank Orders Using Partially Balanced Incomplete Block Designs</i> RALPH G. STRATON .....	555
<i>Estimating Moments of Universe Scores and Associated Standard Errors in Multiple Matrix Sampling for All Item-Scoring Procedures.</i> TEJ N. PANDY AND DAVID M. SHOEMAKER .....	567
<i>The Structure of Domain Hierarchies Found within a Domain Referenced Testing System.</i> GEORGE B. MACREADY .....	583
<i>Some Multiple Range Tests for Variances</i> KENNETH J. LIVY .....	599
<i>Judgmental Bias in the Rating of Attitude Statements.</i> WILLIAM H. BRUVOLD .....	605
<i>Empirical Option Weighting with a Correction for Guessing.</i> RICHARD R. RILEY .....	613
<i>The Concept of Efficiency in Item Analysis.</i> RICHARD J. HOIMANN .....	621
<i>Factor Structure of the McCarthy Scales at Five Age Levels between 2½ and 8½.</i> ALAN S. KAUFMAN .....	641
<i>Possible Sampling Bias in Genetic Studies of Genius.</i> DANIEL P. KEATING .....	657

## COMPUTER PROGRAMS

<i>Univocal Varimax: An Orthogonal Factor Rotation Program for Optimal Simple Structure.</i> DOUGLAS N. JACKSON AND HARVEY A. SKINNER .....	663
---	-----

(Continued on inside front cover)

(Continued from front cover)

<i>The Path Analysis of Complex Recursive Systems.</i> CHARLES F. TURNER	667
<i>IRIS: A Computer-Interactive APL Program for Recovering Simple Orders.</i> THOMAS J. REYNOLDS AND NORMAN CLIFF	671
<i>INTERORD: A Computer-Interactive FORTRAN IV Program for Developing Simple Orders.</i> JERARD F. KEHOE AND NORMAN CLIFF	675
<i>Mapping Individual Logical Processes.</i> FREDERICK O. SMETANA	679
<i>ITANA—III: A FORTRAN IV Program for Multiple-Choice Tests and Item Analysis.</i> BARUKH NEVO, ELI SHOR, AND RACHEL RAMRAZ	683
<i>A Program System for the Estimation of Characteristics of the Test Score Distribution Resulting from Test Items with Given Statistics.</i> LEE L. SCHROEDER	685
<i>A Computer Program to Calculate Adjusted and Unadjusted Interrater Reliabilities for Sets and Subsets of Judges.</i> JOHN F. GREENE, WILLIAM M. MCCOOK, AND FRANCIS X. ARCHAMBAULT	689
<i>A Program for the T-Score Normal Standardizing Transformation.</i> RONALD C. WIMBERLEY	693
<i>Distinguishing Blanks from Zeros in FORTRAN on the IBM 360 Computer.</i> NATHAN JASPEN	697
<i>The Calculation of Correlation Matrices Using Single Subscript Notation.</i> NATHAN JASPEN	701
<i>A Computer Program to Create a Population with Any Desired Centroid and Covariance Matrix.</i> JOHN D. MORRIS	707

(Continued on outside back cover)

This journal is open to: (1) discussions of problems in the field of the measurement of individual differences, (2) reports of research on the development and use of tests and measurements in education, industry, and government, (3) descriptions of testing programs being used for various purposes, and (4) miscellaneous notes pertinent to the measurement field, such as suggestions of new types of items or improved methods of treating test data. Authors are granted permission to have reprints made of their own articles for their own use at their own expense. Manuscripts should be sent in duplicate to Dr. W. Scott Gehman, Box 6907 College Station, Durham, North Carolina 27708. Authors are requested to put tables, footnotes, and abstracts on pages separate from the text and to follow the general directions given in the *Publication Manual of the American Psychological Association (1974 Revision)*. Journal titles should not be abbreviated.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT is published quarterly, one volume per calendar year, at 3121 Cheek Road, Durham, North Carolina 27704. Second class postage paid at Durham, North Carolina and other cities.

Publication charges to authors are as follows: \$30.00 per page of running text; \$40.00 per page of tables, figures, and formulas.

Subscription rate, \$20.00 a year, domestic and foreign. Single copies, \$5.00. Back volumes: Volumes 5 to the present \$20.00 each. See inside back cover for information relative to back issues.

Orders should be sent to EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, Box 6907, College Station, Durham, North Carolina 27708.

Copyright © 1975 by Frederic Kuder



# EDUCATIONAL and PSYCHOLOGICAL MEASUREMENT

**A QUARTERLY JOURNAL DEVOTED TO THE DEVELOPMENT AND  
APPLICATION OF MEASURES OF INDIVIDUAL DIFFERENCES**

<i>Group Size Effects in Employment Testing.</i> JOSEPH M. HILLERY AND STEPHEN S. FUGITA .....	745
<i>Highest Entry Hierarchical Clustering.</i> LOUIS L. MCQUITTY AND VALERIE L. KOCH .....	751
<i>Q Factor Analysis: Applications to Educational Testing and Program Evaluation.</i> F. STEVENS REDBURN .....	767
<i>The Calculation of Reliability from a Split-Plot Factorial Design.</i> ROBERT L. BRENNAN .....	779
<i>Some Comments Concerning the Use of Monotonic Transformations to Remove the Interaction in Two-Factor ANOVA's.</i> SCHUYLER W. HUCK AND CARY O. SUTTON .....	789
<i>Comparing the Variances of Several Treatments with a Control.</i> KENNETH J. LEVY .....	793
<i>Negative Similarities.</i> ULF LUNDBERG AND BERNARD DEVINE .....	797
<i>Scaling Attitude Items: A Comparison of Scalogram Analysis and Order- ing Theory.</i> PETER W. AIRASIAN, GEORGE F. MADAUS, AND ELINOR M. WOODS .....	809
<i>The Differential Formation of Response Sets by Specific Determiners.</i> PHILLIP D. JONES AND GARY G. KAUFMAN .....	821
<i>Structure of Academic Attitudes and Study Habits.</i> S. B. KHAN AND DEN- NIS M. ROBERTS .....	835
<i>A Measure of Reliability Using Qualitative Data.</i> MENI KOSLOWSKY AND HOWARD BAILIT .....	843
<i>Reliable Dimensions for WISC Profiles.</i> ANTHONY J. CONGER AND JUDITH COHEN CONGER .....	847
<i>Paired Comparisons Intransitivity: Trends across Domains of Content and across Groups of Subjects.</i> DARWIN D. HENDEL .....	865
<i>An Analysis of the Meaning of the Question Mark Response Category in Attitude Scales.</i> BERNARD DUBOIS AND JOHN A. BURNS .....	869

## VALIDITY STUDIES OF ACADEMIC ACHIEVEMENT

<i>Section Selection in Multi-Section Courses: Implications for the Valida- tion and Use of Teacher Rating Forms.</i> LES LEVENTHAL, PHILIP C. ABRAMI, RAYMOND P. PERRY, AND LAWRENCE J. BREEN .....	885
<i>Esteem Construct Generality and Academic Performance.</i> C. KENNETH SIMPSON AND DAVID BOYLE .....	897

(Continued on inside front cover)



<i>The Validity of Some Alternative Measures of Achievement Motivation.</i> FRANK B. W. HARPER .....	905
<i>Relationships among Four Measures of Achievement Motivation.</i> THOMAS R. WOTRUBA AND KARL F. PRICE .....	911
<i>Prediction of Persistence and Performance with the Hermans Prestatic Motivation Test.</i> J. OGDEN HAMILTON .....	915
<i>A Preliminary Validation of an Instrument to Measure the Degree of Counselor Restrictive-Nonrestrictive Cognitive Orientation.</i> THOMAS A. SEAY AND F. TERRILL RILEY .....	921
<i>High School Type, Sex, and Socio-economic Factors as Predictors of the Academic Achievement of University Students.</i> JOHN F. McDONALD AND MICHAEL S. MCPHERSON .....	929
<i>Relationships of Selected Nonacademic and Academic Variables to the Grade Point Average of Black Students.</i> SHIH-SUNG WEN AND ROSE E. MCCOY .....	935
<i>Comparative Prediction of First Year Graduate and Professional School Grades in Six Fields.</i> LEONARD L. BAIRD .....	941
<i>The Moderator Effect of Undergraduate Grade Point Average on the Prediction of Success in Graduate Education.</i> ROBERT W. COVERT AND NORMAN M. CHANSKY .....	947
<i>Predictive Validity of SVIB Pharmacist Scales.</i> RICHARD W. JOHNSON, KENNETH W. KIRK, AND RICHARD A. OHVALL .....	951
<i>Use of Selected Factors as Predictors of Success in Completing a Secondary Teacher Preparation Program.</i> FRANK P. BELCASTRO .....	957
<i>The Prediction of Performance in an Educational Psychology Master's Degree Program.</i> ANDREW BEAN .....	963
<i>The Relationship of the Watson-Glaser Critical Thinking Appraisal to Sex and Four Selected Personality Measures for a Sample of Dutch First-Year Psychology Students.</i> JOH. HOOGSTRATEN AND H.H.C.M. CHRISTIAANS .....	969
<i>Convergent and Discriminant Validities of Two Sets of Measures of Spatial Orientation and Visualization.</i> LEWIS PRICE AND JOHN ELIOT .....	975

(Continued on outside back cover)

This journal is open to (1) discussions of problems in the field of the measurement of individual differences, (2) reports of research on the development and use of tests and measurements in education, industry, and government, (3) descriptions of testing programs being used for various purposes, and (4) miscellaneous notes pertinent to the measurement field, such as suggestions of new types of items or improved methods of treating test data. Authors are granted permission to have reprints made of their own articles for their own use at their own expense. Manuscripts should be sent in duplicate to Dr. W. Scott Gehman, Box 6907 College Station, Durham, North Carolina 27708. Authors are requested to put tables, footnotes, and abstracts on pages separate from the text and to follow the general directions given in the *Publication Manual of the American Psychological Association* (1974 Revision). Journal titles should not be abbreviated.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT is published quarterly, one volume per calendar year, at 3121 Cheek Road, Durham, North Carolina 27704. Second class postage paid at Durham, North Carolina and other cities.

Publication charges to authors are as follows: \$30.00 per page of running text, \$40.00 per page of tables, figures, and formulas.

Subscription rate, \$20.00 a year, domestic and foreign. Single copies, \$5.00. Back volumes. Volumes 5 to the present \$20.00 each. See inside back cover for information relative to back issues.

Orders should be sent to EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, Box 6907, College Station, Durham, North Carolina 27708.

EDUCATIONAL and  
PSYCHOLOGICAL

Bureau of Educ. & Psych. Research  
(S. C. E. R. T.)

MEASUREMENT

W. SCOTT GEHMAN, *Editor*  
GERALDINE R. THOMAS, *Managing Editor*  
WILLIAM B. MICHAEL, *Editor, Validity Studies and Computer Programs*  
JOAN J. MICHAEL, *Assistant Editor, Validity Studies and Computer Programs*  
MAX D. ENGELHART, *Book Review Editor*  
LEWIS R. AIKEN, JR., *Assistant Book Review Editor*  
FREDERIC KUDER, *Editor Emeritus*

BOARD OF COOPERATING EDITORS

DOROTHY C. ADKINS, *University of Hawaii*  
LEWIS R. AIKEN, JR., *Guilford College*  
HAROLD P. BECHTOLDT, *The University of Iowa*  
WILLIAM V. CLEMANS, *American Institutes for Research*  
LOUIS D. COHEN, *University of Florida*  
ANTHONY J. CONGER, *Duke University*  
JUNIUS A. DAVIS, *Research Triangle Institute*  
HAROLD A. EDGERTON, *Performance Research, Inc.*  
GENE V. GLASS, *University of Colorado*  
J. P. GUILFORD, *University of Southern California, Los Angeles*  
JOHN A. HORNADAY, *Babson College*  
JOHN E. HORROCKS, *The Ohio State University*  
CYRIL J. HOYT, *University of Minnesota*  
MILTON D. JACOBSON, *University of Virginia*  
JOSEPH C. JOHNSON II, *Jackson State University*  
WILLIAM G. KATZENMEYER, *Duke University*  
ROBERT E. LANA, *Temple University*  
FREDERIC M. LORD, *Educational Testing Service*  
ARDIE LUBIN, *Navy Medical Neuropsychiatric Research Unit, San Diego*

LOUIS L. MCQUITT, *University of Miami, Coral Gables*  
HOWARD G. MILLER, *North Carolina State University at Raleigh*  
ROBERT L. MORGAN, *North Carolina State University at Raleigh*  
HENRY MOUGHAMIAN, *City Colleges of Chicago*  
DAVID NOVAK, *The Neuse Clinic, New Bern, N. C.*  
ELLIS B. PAGE, *University of Connecticut*  
NAMBURY S. RAJU, *Science Research Associates, Inc.*  
BEN H. ROMINE, JR., *University of North Carolina at Charlotte*  
THELMA G. THURSTONE, *University of North Carolina at Chapel Hill*  
WILLARD G. WARRINGTON, *Michigan State University*  
JOHN L. WASIK, *North Carolina State University at Raleigh*  
KINNARD WHITE, *University of North Carolina at Chapel Hill*  
JOHN E. WILLIAMS, *Wake Forest University*  
E. G. WILLIAMSON, *University of Minnesota*

VOLUME THIRTY-FIVE, NUMBER ONE, SPRING 1975

V. 35  
1975

145





## PRESCHOOL RACIAL ATTITUDE MEASURE II<sup>1</sup>

JOHN E. WILLIAMS,<sup>2</sup> DEBORAH L. BEST, DONNA A. BOSWELL,  
LINDA A. MATTSON, AND DEBORAH J. GRAVES

Wake Forest University

The earlier version of the Preschool Racial Attitude Measure (PRAM I) has been found to be a useful measure in attitude development and modification studies of young children. This paper describes the lengthened and otherwise revised version of this procedure—PRAM II.

Standardization data are reported for 252 Caucasian and 140 Negro children, ranging in age from 37 to 85 months (mean = 64 months), who were tested by Caucasian and Negro examiners. Analyses of the racial attitude scores revealed that the measure had good internal consistency ( $r = .80$ ), and satisfactory test-retest reliability ( $r = .55$ , over a one-year interval). It was demonstrated that the test may be divided into two equivalent short-forms, for test-retest purposes. Other findings were that the racial attitude scores were found to vary systematically with race of subject, but not with sex of subject, IQ, or age. Evidence regarding race of examiner effects was inconclusive.

It was concluded that PRAM II provides a reliable index of racial attitudes, and that the same rationale could be employed in the assessment of other attitudes at the preschool level. Theories of racial attitude development are discussed.

WILLIAMS and Roberson (1967) described a method for the

<sup>1</sup> This study was supported by a grant to the first author from the National Institute of Child Health and Human Development (HD-02821). The authors are indebted to the administrators and teachers of the participating schools for their cooperation. The authors are grateful to Beth Norbrey, Kathleen Williams, Elaine Wright, Shirley Colquett, and Shari Fulmer for their assistance as examiners.

<sup>2</sup> Requests for reprints of this paper, for copies of the PRAM II manual and technical report, and for information concerning the loan or purchase of test materials should be sent to John E. Williams, Department of Psychology, Wake Forest University, Winston-Salem, N.C. 27109.

Copyright © 1975 by Frederic Kuder

assessment of racial attitudes in preschool children. In this procedure, subsequently named the Preschool Racial Attitude Measure I (PRAM I), a series of pictures were employed, each of which contained two human figures, one with pinkish-tan skin and blonde hair ("Caucasian"), the other with medium-brown skin and black hair ("Negro"). Each picture was accompanied by a story containing one of six positive evaluative adjectives or one of six negative evaluative adjectives, with the child being asked which of the two persons was the one described in the story. The child, thus, had twelve opportunities to select one of the two figures in response to the adjectives. With one point given for selecting the Caucasian figure in response to a positive adjective, and one point for selecting the Negro figure in response to a negative adjective, the score range was zero to twelve, with low scores indicating a pro-Negro/anti-Caucasian bias, high scores indicating pro-Caucasian/anti-Negro bias, and scores around six indicating no bias.

The foregoing PRAM I procedure has been employed in a number of investigations, several of which are summarized in Table 1. In the table, it can be seen that the investigations have been conducted in a number of different locales, with groups of children differing in age, race, and social class. It will be noted that the mean racial attitude (RA) scores in all groups fell in the upper portion of the score range, indicating a tendency toward pro-Caucasian/anti-Negro attitudes in all groups. The data indicated, however, that this tendency was stronger among Caucasian children than among Negro children. In addition, Vocke's (1971) data suggests that the race of the examiner may have had a slight effect upon the scores obtained.

Several investigators have employed the PRAM I procedure to assess the outcome of studies designed to modify racial attitudes in preschool children. The general findings from these studies indicated that these attitudes were: easily changed by direct behavior modification procedures (Edwards and Williams, 1970; McMurtry and Williams, 1972); changed, but less dramatically, by modifying the children's affective responses to the colors white and black (Williams and Edwards, 1969; McAdoo, J. L., 1970); and unchanged by special curriculum procedures (McAdoo, J. L., 1970; Walker, 1971).

The original rationale of the PRAM I procedure was derived from that of the semantic differential and was based on the assumption that the "semantic space" of the preschool child embraces an evaluative dimension, similar to that previously demonstrated

TABLE 1  
Mean Racial Attitude (RA) Scores in Various Studies Employing the PRAM I Procedure

Investigator(s)	N	Av. Age	Race of E	Social Class	State-Year	Mean RA
<i>Caucasian Groups</i>						
Williams and Roberson (1967)	111	5-4	Cauc.	M	N. C. ('66)	10.3
Williams and Edwards (1969)	84	5-6	Cauc.	M	N. C. ('67)	9.6
Thompson*	27	3-8	Cauc.	(?)	Calif. ('68)	9.1
Bridges*	31	4-8	Cauc.	M	Texas ('69)	9.7
Bridges*	24	6-11	Cauc.	L	Texas ('69)	11.5
Firestone and Feinstein*	16	4-11	Cauc.	(?)	Conn. ('69)	10.0
Keller*	24	5-9	Cauc.	Mixed (L and M)	Ohio ('70)	9.8
Tse*	30	4-4	Cauc.	M	N. Y. ('71)	10.0
				(Mean of Caucasian Means = 10.0)		
<i>Negro Groups</i>						
Vocke (1971)	45	5-5	Cauc.	L	S. C. ('70)	9.5
Vocke (1971)	45	5-5	Negro	L	S. C. ('70)	8.6
McAdoo, H. (1970)	35	5-1	Negro	L	Mich. ('70)	8.7
McAdoo, H. (1970)	43	5-6	Negro	L	Miss. ('70)	8.9
McAdoo, J. (1970)	65	4-6	Negro	L	Mich. ('70)	7.8
					(Mean of Negro Means = 8.7)	

\* Data from unpublished studies.



among older children and adults by Osgood, Suci, and Tannenbaum (1957), DiVesta (1966), and others. This evaluative dimension is conceptualized as being defined at one extreme by a group of non-synonymous words which share a common meaning response of positive evaluation, or "goodness"; and at the other extreme by a second group of non-synonymous words which share a common meaning response of negative evaluation or "badness." In a series of recent studies, it has been demonstrated that the assumption of such an evaluation dimension at the preschool level is tenable (Edwards and Williams, 1970; McMurtry and Williams, 1972; Gordon and Williams, 1973). This makes it feasible to employ the evaluative dimension to develop preschool attitude measures which can be coordinated with traditional adult attitude measures via the rationale of the semantic differential.

The purpose of the present paper was to describe a revised version of the Preschool Racial Attitude Measure, designated PRAM II, and to present initial standardization data for the new procedure. Among other changes, the revision involved a redrawing of the stimulus figures, a doubling of the length of the procedure, the study of possible race of subject and race of examiner effects, and a careful examination of psychometric characteristics.

### *Method<sup>3</sup>*

#### *Materials*

In the PRAM II revision, several changes were made in the test materials. A new set of twenty-four  $8 \times 10$  racial attitude pictures was drawn in order to improve the general artistic quality of the stimulus materials. The skin-colors of the two figures in each picture were the same as in PRAM I but the difference in hair color was removed by drawing both figures with black hair. The two PRAM II figures in each picture were, thus, identical except for the skin-color difference—pinkish-tan vs. medium brown. In the series of 24 pictures, figures of both sexes were employed, and a variety of ages—from young children to "grandparents"—were represented. The figures in the series were drawn in a variety of sitting, standing, and walking positions, with the pictures being otherwise generally ambiguous as to any activities in which the persons represented might be engaged.

<sup>3</sup> The PRAM II materials and procedures are described more fully in the PRAM II manual (Williams, 1971a), available upon request.

The list of six positive and six negative adjectives employed in PRAM I were each doubled by the addition of six more adjectives. The twelve positive adjectives used in PRAM II were: clean, good, kind, nice, pretty, smart; and friendly, happy, healthy, helpful, right, and wonderful. The twelve negative adjectives used were: bad, dirty, mean, naughty, ugly, stupid; and cruel, sad, selfish, sick, unfriendly, and wrong. In both adjective groups, the first six adjectives were the "old" adjectives used in PRAM I, while the second six are the "new" adjectives added in the PRAM II revision. In PRAM II, the old and new adjectives were equally distributed between the first half of the test (called Series A), and the second half of the test (Series B).

PRAM I contained, in addition to the racial attitude items, a series of twelve sex-role items which assessed the child's knowledge of typical sex-stereotyped behaviors, and which provided a control measure of general conceptual development. These same items (see Williams and Roberson, 1967) were incorporated into the PRAM II procedure. For these items, a new series of twelve  $8 \times 10$  sex-role pictures was drawn, each of which displayed a male and female figure of the same general age, and of the same race (half of the pictures represented Caucasians; half, Negroes).

In summary, the materials for the total PRAM II procedure consisted of 36 pictures, 24 of which were used for racial attitude items and 12 for sex-role items. In the standard administration of the procedure, the first item was a sex-role item, followed by two racial attitude items, with this pattern repeated throughout the test.

### *Subjects*

The *basic standardization group* for PRAM II consisted of 272 preschool children from Winston-Salem, North Carolina.<sup>4</sup> The children ranged in age from 37 months to 85 months, with a mean age of 64.9 months (S.D. = 7.64). Half of the children were Caucasian, and half were Negro, with each race group composed of equal numbers of males and females. As described below, half of each race-sex group were tested by Caucasian examiners, and half by Negro examiners. A three dimensional analysis of variance (race of subject  $\times$  sex  $\times$  race of examiner) indicated that the mean chronological age in all groups was equivalent. The principal data analyses reported below were based on this group of subjects.

<sup>4</sup> This group was a further expansion of the 1970-71 standardization group ( $N = 232$ ) described in the PRAM II Technical Report #1. The additional subjects served to correct the sex imbalance in the 1970-71 group.

Data were also available for a supplemental group of 116 Caucasian and 4 Negro preschoolers with a mean chronological age of 62.0, tested primarily by Caucasian examiners. For certain analyses, these subjects were combined with the 272 subjects from the standardization group to produce a *combined group* of 392 subjects with a mean age of 64.0 months, 252 of whom were Caucasian (mean age = 64.2) and 140 of whom were Negro (mean age = 63.6), with both race groups composed of equal numbers of males and females.

Twenty-nine Caucasian and 28 Negro subjects from the standardization group were retested after a one year interval by an examiner of the same race as the one who had done the original testing. The mean chronological age of this *retest group* was 56.8 months at the time of the first testing, and 69.6 months at the time of the second.

### *Procedure*

All examiners were females in their early twenties. Examiners for the standardization study were one Caucasian graduate student, and four undergraduate students, two of whom were Caucasian and two Negro. Two additional undergraduate students, one Caucasian and one Negro, served as examiners in the retest study. The data for the supplementary group was gathered by the above persons with the assistance of two additional undergraduate examiners, one Negro and one Caucasian. All examiners were trained in administering the PRAM II and Peabody Picture Vocabulary Test (PPVT) procedures. Particular care was taken in training examiners not to provide incidental feedback to the children during the PRAM II administration.

The standard procedure for the administration of PRAM II was as follows. The child was taken from his classroom to a private room where he and the examiner were seated at a low table. After some initial conversation to build rapport, the examiner placed the PRAM II picture notebook and answer sheet on the table and said:

"What I have here are some pictures I'd like to show you, and stories to go with each one. I want you to help me by pointing to the person in each picture that the story is about. Here, I'll show you what I mean." The examiner then opened the notebook to the first (sex-role) picture of a little boy and a little girl seated, and read the first story: "Here are two children. One of these children has four dolls with which they like to have tea parties. Which child likes to play with dolls?" After recording the child's

response; the examiner displayed the second picture of two little boys, one Caucasian and one Negro, walking, and read the second story: "Here are two little boys. One of them is a *kind* little boy. Once he saw a kitten fall into a lake and he picked the kitten up to save it from drowning. Which is the *kind* little boy?" After recording the child's response, the examiner proceeded to picture three and story three, etc., until all 36 items (12 sex-role; 24 racial attitude) had been presented. In the standardization study, half the subjects were administered the second half of the test first, in order to study the equivalence of Series A (Items 1-18), and Series B (Items 19-36) of the procedure.

In the standardization group, the PRAM II administration was followed by the administration of the Peabody Picture Vocabulary Test (PPVT), Form B, following standard directions (Dunn, 1965).

The PRAM II racial attitude responses and sex role responses were scored in the following manner. The racial attitude score was determined by counting one point for the selection of the light-skinned figure in response to a positive adjective, and one point for the selection of a dark-skinned figure in response to a negative adjective. The racial attitude total score based on all 24 items thus had a range of 0-24, with high scores indicating a pro-Caucasian/anti-Negro bias, low scores indicating a pro-Negro/anti-Caucasian bias, and mid-range scores (around 12) indicating no bias. In addition to this total score, several pairs of subscores were determined for each subject. Each pair of subscores was based on a division of the subject's responses into two halves, and each subscore thus had a range of 0-12: (1) a first-half (Series A) score, and a second-half (Series B) score; an odd-numbered items score, and an even-numbered items score; an old items score and a new items score; and a positive adjective score and a negative adjective score. The 12 sex-role items were scored by giving one point for each sex-appropriate response, yielding a possible score range of 0-12. The PPVT was scored in standard fashion to yield an IQ score for each subject.

### *Results<sup>5</sup>*

The total racial attitudes scores (RA-T) for the 272 subjects in the standardization group were analyzed by race of subject, race of examiner, and sex of subject. The three-dimensional analysis of

---

<sup>5</sup> Additional detailed results are available in the PRAM II Technical Report #1 (Williams, 1971b), available upon request.



variance performed to assess the effects of these variables revealed significant ( $p < .01$ ) main effects of race of subject and race of examiner, and a nonsignificant main effect of sex. In addition, no interactions were significant. The nature of the two significant main effects can be seen in Table 2 which presents the mean RA-T scores in each of the four race of subject/race of experimenter groups. The race of subject effect is seen in the fact that Caucasian children obtained higher scores than Negro children under both race of examiner conditions, with an overall Caucasian subject mean of 17.02 and an overall Negro subject mean of 14.73. The significant race of examiner effect indicates that both groups of children obtained higher scores with Caucasian examiners than with Negro examiners, with an overall Caucasian examiner mean of 16.97, and an overall Negro examiner mean of 14.78. The suggestion in Table 2 that the race of examiner effect is greater for Caucasian children than for Negro children was not supported by the statistical analysis, since the race of the subject by race of examiner interaction effect was not statistically significant.

Data from the 57 children in the retest group were used to recheck the two main effects just described. As noted earlier, these children were retested after a one year interval by different examiners of the same race as their initial examiners. An analysis of the RA-T scores in this group again revealed a significant ( $p < .01$ ) race of subject effect: Caucasian subject  $\bar{X} = 18.8$ ; Negro subject  $\bar{X} = 15.0$ . When the data was examined by race of examiner, no significant difference was found. An additional source of negative evidence regarding race of examiner effects is found in a study by Best (1972). In this study, each of 60 preschool Caucasian children was administered PRAM II by two examiners. The first examiner gave standard instructions and administered the first half of PRAM

TABLE 2

*Mean Total Racial Attitude Scores for 272 Subjects Classified by Race of Subject and Race of Examiner*

		Race of Examiner		Total
		Caucasian	Negro	
<i>Race of Subject</i>	Caucasian	18.66 ( $N = 68$ )	15.38 ( $N = 68$ )	17.02 ( $N = 136$ )
	Negro	15.28 ( $N = 68$ )	14.18 ( $N = 68$ )	14.73 ( $N = 136$ )
	Total	16.97 ( $N = 136$ )	14.78 ( $N = 136$ )	15.88 ( $N = 272$ )

II. At this point, a second examiner entered the room, replaced the first examiner, and administered the second half of the procedure. One quarter of the subjects were tested by each of the following race of examiner combinations: two Caucasians; two Negroes; Caucasian then Negro; and Negro then Caucasian. The analyses of the data obtained under these conditions provided no evidence of race of examiner effects. For example, the mean RA-T score obtained by the two Caucasian examiners was 17.6, compared with 17.9 for the two Negro examiners. In addition, there was no evidence of a tendency to shift scores up or down when the race of examiner was changed. In sum, the race of examiner effect found in the standardization group was not replicated in either the retest group, or in the Best study. In view of this, it appears that, for the present, the evidence regarding race of examiner effects at the preschool level remains inconclusive.

Returning to the race of subject effect, it was noted above that this effect was replicated in the retest study. These findings were also congruent with the findings from the PRAM I studies summarized earlier in Table 1. Thus, it was concluded that Negro children as a group make lower racial attitude scores than do Caucasian children. We will now examine the distribution of scores in each of these groups more closely by employing the data from the combined group and ignoring the question of race of examiner effects.

Due to the two-choice nature of the PRAM procedure, the binomial distribution provided a convenient way to determine when an individual child was responding in a manner which would be unlikely on a chance basis. With 24 response opportunities, the probability of an unbiased child obtaining a score of 17 or up was only .035; the same probability existed for scores of 7 or down. Thus, scores in the former category (17 up) were taken as evidence of a "definite" pro-Caucasian/anti-Negro bias ( $C+/N-$ ), while scores in the latter category (7 down) reflected a "definite" pro-Negro/anti-Caucasian bias ( $N+/C-$ ). Likewise, scores of 15 and 16, 8 and 9, were taken as evidence of "probable" bias, while the 10-14 mid-range was characterized as "unbiased."

In Table 3 are presented the percentages of the 252 Caucasian children and 140 Negro children from the combined group who obtained scores in each of the foregoing classes. When the observed frequencies in each class were compared with the chance frequencies from the binomial distribution, a significant ( $p < .001$ ) chi square was obtained for each group. There was a significant

TABLE 3

*Percentage of Total Radical Attitude (RA-T) Scores of Preschool Children Falling Into Each of Five Categories*

RA-T Score Range	Category	Caucasian Subjects (N = 252)	Negro Subjects (N = 140)	Chance Expectancy
17-24	Definite C+/N- Bias	60.3%	39.3%	3.3%
15-16	Probable C+/N- Bias	11.5%	12.9%	12.1%
10-14	Non-Biased	23.4%	32.1%	69.2%
8-9	Probable N+/C-	2.8%	5.0%	12.1%
0-7	Definite N+/C- Bias	2.0%	10.7%	3.3%

tendency toward high (pro-Caucasian/anti-Negro) scores among both Caucasian and Negro children. Perhaps the most dramatic evidence seen in Table 3 was the high degree of definite C+/N- bias (17 up) in both subject groups, being found in approximately 6 out of 10 Caucasian children and 4 out of 10 Negro children. At the other extreme, evidence of definite N+/C- bias (7 down) was found in only 1 out of 10 Negro children and in only 1 out of 50 Caucasian children.

#### *Relationship of Racial Attitude to Other Subject Variables*

The lack of relationship of RA-T scores to sex of subject was demonstrated by the nonsignificant sex effect in the analysis of variance described above. The possible relationships between RA-T scores and other subject variables were explored by means of product-moment correlation coefficients computed among the variables of RA-T scores, age, PPVT-IQ scores, and Sex Role scores. These correlations are summarized in Table 4 where it can be seen that the RA-T scores were independent of both chronological age and PPVT-IQ. On the other hand, the Sex Role score, with which RA-T shows only a slight positive relationship, is significantly associated with both chronological age and PPVT-IQ. These findings appear to have important theoretical implications regarding the origin of racial attitudes which will be discussed later in the paper.

#### *Internal Consistency*

The internal consistency of the 24-item RA-T scale was examined using data from the 392 subjects in the combined group. This was done by comparing the subjects' responses to two halves of the scale, with the data divided in several different ways: odd items *vs.* even items; "old" adjectives *vs.* "new" adjectives; positive ad-



jectives *vs.* negative adjectives; and first half (Series A) *vs.* second half (Series B). In each of these comparisons a product moment correlation coefficient was computed between scores on the two 12-item halves, and the Spearman-Brown correction for doubled length was then employed to estimate the internal consistency of the total 24-item scale.

The results of these comparisons are summarized in Table 5. These findings indicated that the racial attitude scale possessed a high degree of homogeneity. The Spearman-Brown estimates for the usual "split-half" comparisons (odd-even; first half-second half) indicated that the internal consistency "reliability" of RA-T scores was of the order of .80.

The findings for the old-item *vs.* new-item comparisons provided satisfactory evidence that the twelve items added in the PRAM II revision were measuring the same thing as the old items from PRAM I. The results for the positive adjective *vs.* negative adjective comparison indicated that scores on these two sub-scales were substantially related, indicating that children who chose light-skinned figures in response to positive adjectives also tended to choose dark-skinned figures in response to negative adjectives, and vice-versa. Thus, the positive and negative items appeared to be assessing different aspects of the same trait.

The results of the Series A *vs.* Series B comparison were of particular interest since the series had been designed to provide alternate short forms of the procedure. The high correlation between A and B scores (.71), the virtually identical means ( $A = 8.20$ ;  $B =$

TABLE 4

*Intercorrelations among Total Racial Attitude Scores, Age, PPVT-IQ, and Sex Role Scores for 252 Caucasian and 140 Negro Preschoolers*

	Subject Groups	Chronological Age	PPVT IQ	Sex Role Score
Racial Attitude Scores	Cauc. Ss	.11	.00	.25*
	Negro Ss	.03	.06	.02
	All Ss	.09	.15	.20*
Chronological Age	Cauc. Ss		.04	.41*
	Negro Ss		.05	.38*
	All Ss		.10	.38*
PPVT-IQ	Cauc. Ss			.12
	Negro Ss			.33*
	All Ss			.33*

*Note.*—Correlations involving PPVT-IQ are based on Caucasian  $N = 136$ , Negro  $N = 132$ , Total  $N = 268$ .

\* $p < .01$ .

TABLE 5

*Internal Consistency Measures for Racial Attitude Scale: Means, Correlation Coefficients (r), and Spearman-Brown Estimates (r SB)*

<i>Odd-Numbered Items vs. Even Numbered Items</i>				
	$\bar{X}$ Odd	$\bar{X}$ Even	r	r SB
Cauc. Ss (N = 252)	8.86	8.55	.61	.76
Negro Ss (N = 140)	7.38	7.31	.76	.86
Total Ss (N = 392)	8.33	8.11	.69	.81
<i>Old Items vs. New Items</i>				
	$\bar{X}$ Old	$\bar{X}$ New	r	r SB
Cauc. Ss	9.23	8.18	.70	.78
Negro Ss	7.64	7.04	.68	.81
Total Ss	8.67	7.77	.70	.82
<i>Positive Items vs. Negative Items</i>				
	$\bar{X}$ Positive	$\bar{X}$ Negative	r	r SB
Cauc. Ss	8.91	8.50	.53	.69
Negro Ss	7.56	7.11	.68	.81
Total Ss	8.45	8.00	.61	.76
<i>First Half (Series A) Items vs. Second Half (Series B) Items</i>				
	$\bar{X}$ Series A	$\bar{X}$ Series B	r	r SB
Cauc. Ss	8.62	8.79	.65	.79
Negro Ss	7.45	7.24	.75	.86
Total Ss	8.20	8.24	.71	.83

8.24) and standard deviations ( $A = 2.74$ ;  $B = 2.79$ ), indicated that the two scales could be considered as equivalent 12 item short forms of PRAM II.

### *Stability*

The stability of the RA-T scores across a one-year interval (12.8 months) was assessed using the 57 subjects from the retest group. At the time of first testing, these subjects had a mean age of 56.8 months; at the second testing, these subjects had a mean age of 69.6 months. The mean RA-T score at the first testing was 15.30 while the mean at the second testing was 16.93, a statistically significant ( $p < .05$ ) increase of 1.63 points. This finding suggests that there may have been a slight positive practice effect for the RA-T scores. It does not seem likely that the increase was attributable to the fact that the children were a year older, since RA-T scores were found not to be correlated with age, as noted above.

Three scores (Series A, Series B, and Total) from the first administration were each correlated with the same three scores from the second administration, with statistically significant coefficients obtained in all instances. The correlation of .55 between total scores

at the two administrations provided the best available estimate of test-retest reliability, although the relative youth of the subjects suggested that this may be a minimum estimate. As is usual, this value was lower than the estimated internal consistency of .80, noted above.

### *Discussion*

The results of the studies just described have led to the conclusion that the PRAM II revision was generally successful and that PRAM II can be viewed as a lengthened and otherwise improved version of PRAM I which provides a psychometrically-sound method for the assessment of racial bias in pre-literate children. While the results of the studies regarding possible race of examiner effects were inconclusive, the evidence regarding race of subject effects was clear and consistent: the racial attitude scores of Negro preschoolers averaged two to three points lower than the scores of Caucasian preschoolers. This evidence of difference in mean scores should not obscure the fact that in *both* groups, pro-Caucasian/anti-Negro ( $C+/N-$ ) bias was much more evident than was the reverse pro-Negro/anti-Caucasian ( $N+/C-$ ) bias.

Evidence such as the foregoing is often interpreted in accord with the "normative" theory of prejudice which would attribute the  $C+/N-$  bias among preschoolers to their having acquired the anti-Negro prejudices of the Caucasian-dominated larger society. This acquisition process is presumed to be a gradual one resulting from the child's repeated contacts with such prejudice through the preschool years. It would seem to follow from the normative theory that whenever children are acquiring any general concept/attitude "from the culture", it would be expected that: (1) progressively older groups of children will show progressively more evidence of the concept/attitude; and (2) brighter children will show more evidence of the concept/attitude than will duller children. Such, in fact, were the findings of the present study for the sex-role scores which reflected the child's knowledge of sex-appropriate behaviors. Regarding racial attitude scores, however, neither of these conditions were met: racial attitude scores were not correlated either with chronological age or with PPVT-IQ scores. Thus, high racial attitude scores were equally evident among brighter and duller children, and older and younger children. These findings are inconsistent with the requirements of the general normative theory.

There remain at least two other plausible theories concerning



the origin of the reliable individual differences in attitude scores. It is possible that the children's attitude scores are related to individual differences in familial variables, e.g., attitudinal and/or cognitive characteristics of parents. A second possibility is that the children's attitudes toward light and dark persons are influenced at least in part by individual differences in early learning experiences with light and darkness ("fear of the dark") which may be essentially independent of familial influences. In any event, it is clear that additional research in the areas of familial influence and/or early experiences will be needed to clarify the origins of the attitudes being assessed by PRAM II.

The preceding discussion raises the question as to whether the attitude trait measured by PRAM II should be designated as *racial* attitude, as opposed to "skin-color" attitude. This question is not easily answered. Part of the difficulty is the lack of systematic knowledge as to what "race" means to preschool children. Some evidence indicates that, at this age, skin-color is the most salient feature associated with race; and it has been demonstrated that preschoolers show little hesitancy in identifying persons as "white," "Negro," "colored," etc., when skin-color is the only basis for discrimination (Williams and Roberson, 1967). Hence, attitude toward skin-color and attitude toward "race" may be virtually synonymous at this age level. While the attitudes assessed may or may not be "racial" in their origins, it seems clear that they are "racial" in their implications, and it is this latter usage which seems to warrant the designation of the PRAM II scores as measures of racial attitude.

It would appear that the PRAM II procedure represents a substantial advance in attitude assessment procedures for preschool children and should facilitate the study of many interesting and important questions dealing with the origins, development, and modifiability of racial attitudes in young children. Regarding the latter, three studies employing PRAM II have already been conducted (Graves, 1973; Shanahan, 1972; Yancey, 1972), and the availability of alternate short forms of the PRAM II procedure should be of benefit to other experimenters who wish to conduct attitude change studies employing a pre-post design. The general evaluation dimension rationale on which PRAM II is based can also be used to assess other types of attitudes in young children. For example, the rationale has been employed in a number of studies dealing with children's attitudes toward the colors black and white (Renninger and Williams, 1966; Williams and Roberson,

1967; Williams and Edwards, 1969; Skinto, 1969; Williams and Rousseau, 1971; Figura, 1971; Vocke, 1971; Gordon and Williams, 1973). In still another research area, the evaluative dimension rationale is currently being employed by the authors in pilot studies aimed at assessing attitudes of preschool children toward male and female persons.

## REFERENCES

- Best, D. L. *Race of examiner effects on the racial attitude responses of preschool children*. Master's Thesis, Wake Forest University, 1972.
- DiVesta, F. J. A developmental study of the semantic structures of children. *Journal of Verbal Learning and Verbal Behavior*, 1966, 5, 249-259.
- Dunn, L. M. *Expanded manual for the Peabody Picture Vocabulary Test*. Circle Pines, Minn.: American Guidance Service, Inc., 1965.
- Edwards, C. D. and Williams, J. E. Generalization between evaluative words associated with racial figures in preschool children. *Journal of Experimental Research in Personality*, 1970, 4, 144-155.
- Figura, A. L. *The effect of peer interaction on the self-concept of Negro children*. Master's Thesis, DePaul University, 1971.
- Gordon, L. H. and Williams, J. E. Secondary factors in the affective meaning system of the preschool child. *Developmental Psychology*, 1973, 8, 25-34.
- Graves, D. J. *Modification of racial attitudes in Caucasian children as a function of curriculum and race of teachers*. Honor's Thesis, Wake Forest University, 1973.
- McAdoo, H. P. *Racial attitudes and self-concepts of Black preschool children*. Doctoral Dissertation, University of Michigan, 1970.
- McAdoo, J. L. *An exploratory study of racial attitude change in Black preschool children using differential treatments*. Doctoral Dissertation, University of Michigan, 1970.
- McMurtry, C. A. and Williams, J. E. The evaluation dimension of the affective meaning system of the preschool child. *Developmental Psychology*, 1972, 6, 238-246.
- Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. *The measurement of meaning*. Urbana: University of Illinois Press, 1957.
- Renninger, C. A. and Williams, J. E. Black-white color connotations and race awareness in preschool children. *Perceptual and Motor Skills*, 1966, 22, 771-785.
- Shanahan, J. K. *The effects of modifying black-white concept attitudes of Black and White first grade subjects upon two measures of racial attitudes*. Doctoral Dissertation, University of Washington, 1972.
- Skinto, S. M. *Racial awareness in Negro and Caucasian elementary school children*. Master's Thesis, West Virginia University, 1969.

- Vocke, J. M. *Measuring racial attitudes in preschool Negro children*. Master's Thesis, University of South Carolina, 1971.
- Walker, P. A. *The effects of hearing selected children's stories that portray Blacks in a favorable manner on the racial attitudes of groups of Black and White kindergarten children*. Doctoral Dissertation, University of Kentucky, 1971.
- Williams, J. E. *Preschool Racial Attitude Measure II (PRAM II): general information and manual of directions*. Department of Psychology, Wake Forest University, Winston-Salem, North Carolina, 1971a.
- Williams, J. E. *Preschool Racial Attitude Measure II (PRAM II): technical report #1: 1970-71 standardization study*. Department of Psychology, Wake Forest University, Winston-Salem, North Carolina, 1971b.
- Williams, J. E. and Edwards, C. D. An exploratory study of the modification of color concepts and racial attitudes in preschool children. *Child Development*, 1969, 40, 737-750.
- Williams, J. E. and Roberson, J. K. A method of assessing racial attitudes in preschool children. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1967, 27, 671-689.
- Williams, J. E. and Rousseau, C. A. Evaluation and identification responses of Negro preschoolers to the colors black and white. *Perceptual and Motor Skills*, 1971, 33, 587-599.
- Yancey, A. V. *A study of racial attitudes in white first grade children*. Master's Thesis, Pennsylvania State University, 1972.



## STRATEGIC USE OF RANDOM SUBSAMPLE REPLICATION AND A COEFFICIENT OF FACTOR REPLICABILITY<sup>1</sup>

WILLIAM G. KATZENMEYER

Duke University

A. JACKSON STENNER

IBEX Corp.

The problem of demonstrating replicability of factor structure across random samples is addressed. Procedures are outlined which combine the use of random subsample replication strategies with the correlations between factor score estimates across replicate pairs to generate a coefficient of replicability and confidence intervals associated with the coefficient. Data from the national norming sample of the Self Observation Scales are used in the illustrative example.

THE importance of demonstrating the replicability of factor structures has been the subject of discussion by Royce (1966) and Nesselroade and Baltes (1970). Nesselroade and Baltes argue: "It should be emphasized that, in general, comparative factor analysis remains in the dilemma of being at best a descriptive technique until more systematic information on factor similarity coefficients is available." As early as 1947, Thurstone (1947) suggested that generalizability of factors is a major objective of factor analytic studies. Cattell (1966) stated: "The interpretation of factors, i.e., the inferring of their natures as scientific determiners, is closely tied to the problems of pattern matching and identification. (For interpreting a factor that has appeared in only one study would not be profitable as a rule.)" While in agreement with Cattell

<sup>1</sup> Special appreciation is extended to Richard Gorsuch and John Nesselroade for their comments on an earlier draft of this paper.

on this issue, many factor analysts find themselves in the position of interpreting such factors.

There is consensus on the desirability and importance of demonstrating factor replicability and invariance, and there are several indices of quantitative invariance available (Burt, 1948; Cattell, 1944; Kaiser, 1960; Tucker, 1951). This consensus has led to the development of several sophisticated rotational techniques designed to produce a high degree of similarity between factor loading patterns from different sets of data (Browne and Kristoff, 1969; Cattell and Cattell, 1955; Eyferth and Sixtl, 1965; Fischer and Roppert, 1964; Kristoff, 1964; Meredith, 1964; Taylor, 1967; Tucker, 1951). Nesselroade and Baltes (1970) have pointed out two problems characteristic of the practice of rotating factor structures to maximum similarity: (a) Such procedures abandon the original rotation criteria, and (b) produce similarity coefficients with random data which approximate the magnitude of similarity coefficients found in studies using observed data. A third problem of these procedures and of several indices of quantitative invariance that have been proposed (Burt, 1948; Cattell, 1944; Kaiser, 1960; Tucker, 1951) is that adequate distributional data are not available. A final problem of procedures which deal with the factor pattern or factor structure matrices will be considered later in this study. Pinneau and Newhouse (1964) reviewed the commonly available indices of factor invariance and offered the coefficient of invariance as an alternative index of the extent to which two factors approach congruence. The coefficient of invariance operates on the factor scores rather than the factor pattern or factor structure matrices. It is the correlation between the estimated factor scores of subjects on two independently derived factors thought to be matched. This correlation is a useful measure of invariance of factor structures obtained from the analysis of a common set of variables on different sets of subjects. Throughout this paper the term "replicability" will be used to refer to the reappearance of the same factors across random samples, while the term "invariance" will refer to the reappearance of the same factors when subjects have been systematically selected.

A promising element of research methodology relevant to demonstrating replicability and invariance of factors is found in the random subsample replication strategies. An excellent discussion of these strategies is provided by Finifter (1972). Finifter suggests the following definition for the procedures, which he labels "random subsample replication (RSSR)":

A total body of data is constituted from, or subsequent to its collection, divided into two or more independent random subsamples. Each subsample is a replicate of a particular sampling design. A distribution of outcomes for a parameter being estimated is generated by applying a common analysis procedure to each subsample. The difference observed among the subsample results are then analyzed to obtain an improved (namely, less variable or less biased) estimate of the parameter, as well as, a confidence assessment for that estimate.

The RSSR strategies are particularly well suited to situations in which "statistical tests" have not been developed or are based on shaky theoretical formulations. Finifter argues that, even in cases where adequate theory for a "statistical test" exists, RSSR strategies provide a closer approximation to the ultimate goal of actual replication than do derived distribution theories (Finifter, 1972). Furthermore, when the issue transcends statistical significance and becomes one of practical utility, the family of RSSR strategies takes on increased attractiveness. The problem of demonstrating factor invariance, crucial to establishing the validity of generalizations from factor analytically derived scales, is well suited to the application of RSSR methodology.

Horst (1966) has criticized RSSR strategy in multivariate research. Horst states:

We know that, other things being equal, the more cases we have, the more stable and reliable our results will be. Therefore, for the purposes of both application and generalization, our procedures must be developed on the largest sample available. If we develop a procedure and then cross-validate it, we have ipso facto not developed the best procedure possible from the available data.

This is, of course, not a criticism of RSSR strategies themselves but is a criticism of their improper use. When dealing with constructs derived through factor analytic methods, failure to demonstrate factor replicability and invariance leaves doubts about the validity of such inferences as might be made from the data. A peculiarity of factor analysis is that an increase in sample size does not necessarily lead to an increase in the replicability of the factor structure. Since every rotation of a factor matrix is an arbitrary transformation, the factor structure derived from 10,000 cases involving random responses or a structure derived from 100 such observations is equally nonreplicable in a second random sample. Further, the work of Nesselrode and Baltes (1970) and Nesselrode, Baltes, and Labouvie (1971) demonstrates that factor struc-

SEARCH  
6.3.81



tures derived from random data can be rotated into positions which yield very high coefficients of similarity across several factors. Needed is a procedure which (a) provides an estimate of the agreement among factor structures (invariance) without altering the original rotational criteria, (b) provides an estimate of the confidence limits around this estimate, (c) includes all sources of variability in its estimate of invariance,<sup>2</sup> and (d) bases the final estimate of the factor structure on all of the available data.

The approach implemented by the authors to meet these requirements combines the use of random subsample replication strategies with factor analytic procedures. Correlations between the factor score estimates on matched factors for each replicate pair are used to generate a coefficient of replicability which provides both an estimate of factor replicability and confidence limits associated with the estimate.

### *Method*

The data used in this study were obtained from the national norming and validation of the Primary Level of the Self Observation Scales (SOS)<sup>3</sup> (Stenner and Katzenmeyer, 1973). The subjects of the study included first, second, and third graders who responded to the 50 items of the (SOS) Primary Form A during April and May 1973.

A sample of 6,300 cases was divided into four random subsamples (replicates) of 1,575 cases each. The steps involved in the separate factor analyses of each random subsample are described in 1 through 5. Steps 6-9 describe the procedures involved in computing the coefficients of replicability.

1. A matrix of phi coefficients was computed. When a missing datum was encountered, the mean value for that variable was inserted. The percentage of missing data was less than 4%.

2. Squared multiple correlations were entered as initial communality estimates. Iteration for communalities proceeded until the maximum absolute deviation between iterations dropped below .001.

<sup>2</sup> Similarity coefficients which deal with the factor structure or factor pattern matrices do not include variance due to the indeterminacy associated with estimating factor scores when the common factor model is employed.

<sup>3</sup> The Primary Level of the Self Observation Scales (SOS) is a direct self-report, group-administered instrument comprised of 45 items (Forms A and B). The SOS (primary level) measures five dimensions of children's affective behavior: (1) Self Acceptance, (2) Social Maturity, (3) School Affiliation, (4) Self Security, and (5) Achievement Motivation. The first four scales were factor analytically derived. The fifth was developed using discriminant analysis.

3. A rotation to the varimax criterion was performed.
4. The orthogonal varimax solution was rotated to maximum oblique simple structure, using the Maxplane criterion (hyperplane width .15).
5. The matrix of loadings of the variables on the factors  $V(fe)$  was computed using  $V_{fe} = R_v^{-1}V_{fs}$ , where  $R_v^{-1}$  is the inverse of the matrix of correlations among the variables and  $V_{fs}$  is the oblique factor structure (matrix of correlations of factors and variables).
6. Scores on each variable (question) for the total group (6,300 cases) were converted to  $z$  scores and factor score estimates (least squares regression estimates) were computed for each subject, on each factor, using the four  $V_{fe}$ 's. Since four factors were identified in each of the rotations, this procedure resulted in 16 factor score estimates for each subject: four factor score estimates on each of four factors.
7. Correlations between the estimated and true factor scores were computed (multiple correlation of the estimated scores with the 45 variables of the data matrix, which is also the standard deviation of the estimated factor scores).
8. Correlation coefficients between factor score estimates from each replicate pair (six pairs) were computed. This procedure produced six estimates of the coefficient of replicability for each factor.
9. Coefficients of replicability and confidence intervals associated with these coefficients were obtained in the following manner: Fisher's  $r$  to  $z$  transformation was performed with each of the six coefficients of invariance obtained for each factor. The means and standard deviations of Fisher  $z$  values were obtained and confidence intervals computed ( $p < .05$ ,  $p < .01$ ). The  $r$  equivalents of the mean Fisher  $z$  value and of the 95 and 99% confidence limit  $z$  values were computed.

### Results

Table 1 presents the eigenvalues (squared multiple  $R$ 's in the diagonal) of the correlation matrices derived from the four random subsamples.

The eigenvalues are consistent across replications; the choice of four factors seems appropriate whether by Guttman's middle bound, the Scree test, or interpretability of factors.

Table 2 presents the percentage of variables in hyperplanes of varying widths after completion of the Maxplane rotation.

TABLE 1

*First Five Eigenvalues for the Four Replication Samples*

	Eigenvalue				
	1	2	3	4	5
Subsample					
I	4.62	2.74	1.52	1.03	0.49
II	4.36	2.88	1.57	1.07	0.49
III	4.42	2.93	1.50	1.06	0.46
IV	4.54	2.64	1.62	1.04	0.44

Cattell (1966) suggests as a "rule of thumb" that in most domains from 55 to 85% of the variables in the hyperplane is indicative of adequate simple structure.<sup>4</sup>

Table 3 presents the correlations between factor score estimates on matched factors for each replicate pair. The 1-2 line presents the correlations between the factor score estimates obtained from analysis of sample (replicate) one and the factor score estimates obtained from replicate two.

Table 4 presents the coefficients of replicability for each of the factors, together with the upper and lower limits associated with the 95 and 99% confidence intervals. It is worth noting that the standard error of the coefficient of replicability includes error variance from at least two sources: (a) sampling error resulting from the selection of the random subsamples, and (b) deviations of the factor score estimates from the true factor scores. Guttman

<sup>4</sup> This criterion is, however, a function of the number of factors extracted. Rotation of these same data for 14 factors yielded 80% of the variables in the hyperplane. Another study by the authors (in press) revealed hyperplane counts of 90% in the .15 hyperplane when rotating for seven factors with random data.

TABLE 2

*Percent Variables in Hyperplane for Varying Widths in Each of the Four Replication Samples*

	Hyperplane Width			
	.05	.10	.15	.20
Subsample				
I	33.3	51.7	68.3	72.8
II	36.1	54.4	70.0	72.8
III	28.3	52.2	69.4	72.8
IV	33.9	54.4	68.9	72.2



TABLE 3

*Correlations between Factor Score Estimates on Each Replicate Pair*

Replicate Pair	Self Acceptance	Social Maturity	Self Security	School Affiliation
1-2	.97	.99	.96	.99
1-3	.94	.99	.98	.99
1-4	.97	.97	.99	.99
2-3	.98	.99	.94	.99
2-4	.96	.96	.97	.99
3-4	.91	.95	.97	.98

(1955) has observed that, in common factor analysis, the regression estimates may depart considerably from the true scores and has suggested methods for computing the correlation between the estimates and the true factor scores. Guttman has shown that the theoretical minimum correlation between the alternative factor score estimates drops rapidly as the correlation between the true factor scores and the regression estimates becomes lower.<sup>5</sup> Knowing the theoretical maximum error is, however, not very useful to the researcher, except to generate temperance in the interpretation of his results. The coefficient of replicability provides an estimate of the maximum loss due to the difference between the true factor scores and the regression estimates of these scores. The standard coefficient of invariance, reflecting both the sampling error and the error due to imperfect regression estimates, allows us to set an upper bound on the amount of error that is associated with the difference between the true factor scores and the factor score estimates. It is suggested that  $(1 - r_r^2)$ , where  $r_r$  is the coefficient of replicability

<sup>5</sup> For example, if the correlation between true scores and regression estimates drops to .90, the maximally different set of scores would have a correlation of only .62 with the original estimates.

TABLE 4

*Coefficients of Replicability and Their Associated Confidence Limits*

	$r_r$	Lower Limits		Upper Limits	
		$p < .01$	$p < .05$	$p < .05$	$p < .01$
Self Acceptance	.96	.84	.89	.99	.99
Social Maturity	.98	.89	.92	.99	.99
Self Security	.97	.86	.90	.99	.99
School Affiliation	.99	.96	.97	.99	.99

as defined above) is the total error variance across factor pairs and constitutes the upper limit of the error due to regression estimates.

In the example of this paper the correlations between the regression estimates and the true factor scores are .90 for Self Acceptance, .91 for Social Maturity, .88 for Self Security, and .90 for School Affiliation. Reference to the table of Guttman (1955) reveals that the maximally different sets of factor scores will have correlations in the low sixties with these estimates. Such correlations (low 60's) would be unacceptable and would have serious effects on reliability and usability of the factor scores. The coefficient replicability with its associated confidence limits reveals that the total error, both of regression and of sampling, does not exceed  $(1 - r_r^2)\%$  of the variance. In the case of the factor labeled "Self Acceptance," the best estimate is that  $(1 - .96^2)$  or 7.55% of the total variability is accounted for by the difference between the samples and differences between true factor scores and regression estimates. It can be said with 99% confidence that the between factor variance from all sources does not exceed  $(1 - .84^2)$  or 29.1% of the total variability. It may be observed (99% confidence) that for the factor labeled "School Affiliation" the between factor variance will not exceed  $(1 - .96^2)$  or 7.1% of the total variance. This finding, when compared with the correlation of .90 between the true factor scores and the regression estimates, suggests that, for the normal range of scores, the loss due to using regression estimates of the true score may not be as great as has previously been believed. While not questioning Guttman's derivation, it may be suggested that the deviations of regression estimates from true factor scores may be systematic rather than random. A procedure for partitioning the unexplained variance between matched factors will be presented in a later article.

A clear conclusion of this study is that the Primary Level (Form A) of the Self Observation Scales has a factor structure that can be replicated on independent subsamples of children for which the instrument was designed, and that the level of stability of factors derived from various random subsamples is high. Derived from a sample comprised of children of different races, ages, socio-economic advantage, and geographic regions, the results of this analysis increase the confidence with which the obtained factor structure can be utilized.

The general approach of combining RSSR strategies with factor analytic procedures and computing a coefficient of replicability yields a more stringent and informative test of invariance than do

the approaches which rotate two or more structures to maximum congruence. In addition, the original rotational criteria are not violated in the process of trying to manufacture congruence (order can often be imposed on a chaotic system, but if it is forced, the resultant similarity between the obtained "order" and nature may be purely coincidental).

In combination, RSSR and the coefficient of replicability can contribute significantly to the confidence with which factors are identified and reported. One result of applying this strategy across random subsamples may be a substantial reduction in the number of reported factor analytic studies.

### *Summary*

1. Demonstration of factor stability (invariance) is prerequisite to the legitimate use of such factors in scientific inquiry.

2. Methods for obtaining and estimating factor invariance through post hoc rotational techniques aimed at producing a high degree of configurational similarity between factor loading patterns from various sets of data (e.g., Browne and Kristoff, 1969) present several problems.

a. The original rotation criteria are abandoned.

b. Similarity coefficients derived from the use of these techniques with random data capitalize on chance elements and often produce factor similarity coefficients in the same range as have been reported with real data.

c. Many rotational schemes which operate on factor matrices ignore the error introduced in the common factor model by differences between the true factor scores and the regression estimates of the factor scores.

3. The coefficient of replicability reflects both error due to sampling and error due to the indeterminacy of the common factor model.

4. Random subsample replication techniques provide a method for computing a series of coefficients of replicability across paired random subsamples.

5. The coefficient of replicability provides a method for estimating stability across any number of replications, stating confidence limits concerning the replicability of each factor.

6. The proportion of error introduced by the difference between regression estimates of the factor scores and the true factor scores will not be more than 1 minus the square of the coefficient of repli-



cability. Confidence limits for the true value of  $r$ , are easily obtainable.

7. The correlation between true factor scores and regression estimates of these scores does not set an upper limit on the coefficient of factor replicability.

8. The factor structure of the Primary Self Observation Scales is replicable over random subsamples.

9. RSSR strategies are appropriate to the demonstration of factor invariance in most factor analytic studies. An adequate sample of data for factor analysis is usually adequate for one of the RSSR strategies.

10. The routine use of RSSR strategies and reporting of both standard coefficients of replicability and invariance and their confidence limits would greatly enhance the ability to make inferences from such studies and would probably improve the average quality of factor analytic studies reported in the literature.

## REFERENCES

- Browne, M. and Kristof, W. On the oblique rotation of a factor matrix to a specified pattern. *Psychometrika*, 1969, 34, 237-248.
- Burt, C. The factorial study of temperament traits. *British Journal of Psychology* (Statistical Section), 1948, 1, 178-203.
- Cattell, R. B. The meaning and strategic use of factor analysis. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology*. Chicago: Rand-McNally, 1966.
- Cattell, R. B. and Cattell, A. K. S. Factor rotation for proportional profiles: Analytical solution and an example. *British Journal of Statistical Psychology*, 1955, 8, 83-91.
- Eyferth, K. and Sixtl, F. Bemerkungen zu einem Verfahren zur maximalen Annäherung zweier Faktorenstrukturen aneinander. *Archiv für die Gesamte Psychologie*, 1965, 117, 131-138.
- Finifter, B. M. The generation of confidence: Evaluating research findings by random subsample replication. In H. L. Costner (Ed.), *Sociological methodology*. London: Jossey-Bass Inc., 1972.
- Fischer, G. H. and Roppert, J. Bemerkungen zu einem Verfahren der Transformationsanalyse. *Archiv für die Gesamte Psychologie*, 1964, 116, 98-100.
- Guttman, L. The determinacy of factor score matrices with implications for five other basic problems of common factor theory. *British Journal of Statistical Psychology*, 1955, 8, 65.
- Horst, P. An overview of the essentials of multivariate analysis methods. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology*. Chicago: Rand-McNally, 1966.
- Kaiser, H. F. Relating factors between studies based upon different individuals. Unpublished manuscript, 1960.
- Meredith, W. Rotation to achieve factorial invariance. *Psychometrika*, 1964, 29, 187-206.

- Nesselroade, J. R. and Baltes, P. B. On a dilemma of comparative factor analysis: A study of factor matching based on random data. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1970, 30, 935-948.
- Nesselroade, J. R., Baltes, P. B., and Labouvie, E. W. Evaluating factor invariance in oblique space: Baseline data generated from random numbers. *Multivariate Behavioral Research*, 1971, 6, 233-241.
- Pinneau, S. R. and Newhouse, A. Measures of invariance and comparability in factor analysis for fixed variables. *Psychometrika*, 1963, 29, 271-281.
- Royce, J. R. Concepts generated in comparative and physiological psychological observations. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology*. Chicago: Rand-McNally, 1966.
- Taylor, P. A. The use of factor models in curriculum evaluation: A mathematical model relating two factor structures. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1967, 27, 305-321.
- Thurstone, L. L. *Multiple factor analysis*. Chicago: University of Chicago Press, 1947.
- Tucker, L. R. A method for synthesis of factor analysis studies. *Personnel Research Section Report No. 984*. Washington, D. C.: Department of the Army, 1951.





## DOMAIN VALIDITY AND GENERALIZABILITY<sup>1</sup>

HENRY F. KAISER

University of California, Berkeley, and  
U. S. Coast Guard Academy

WILLIAM B. MICHAEL

University of Southern California

An alternative derivation of Tryon's basic formula for the coefficient of domain validity or the coefficient of generalizability developed by Cronbach, Rajaratnam, and Gleser is provided. This derivation, which is also the generalized Kuder-Richardson coefficient, requires a relatively minimal number of assumptions compared with that in previously proposed approaches.

THIS note provides an alternative derivation of the basic formula for Tryon's (1957) coefficient of domain validity or the coefficient of generalizability offered by Cronbach, Rajaratnam, and Gleser (1963). It appears that the writers' treatment affords a more nearly assumption-free development for these coefficients (which are also the generalized Kuder-Richardson coefficient) than that which has been proposed previously.

The most fundamental notion in statistics is that of making inductive inferences about the characteristics of a population of individuals on the basis of observations on a sample of individuals from the population. Similarly, but not usually made so explicitly, the fundamental problem of psychometrics is to study the nature of a domain or universe of variables on the basis of observations on a sample (or selection) of variables from the domain or universe. (Tryon's "domain" is Guttman's or Cronbach's "universe.") Thus, if there exist  $n$  individuals and  $p$  variables, statistical inference is

<sup>1</sup> The research reported in this note was supported in part by the Office of Computing Activities, National Science Foundation.

Copyright © 1975 by Frederic Kuder

concerned with what happens as  $n$  becomes very large, whereas psychometric inference is concerned with what happens as  $p$  becomes very large. It is in the context of psychometric inferences that the confidence one may have in the observed score on a test is to be considered.

The observed score,  $\hat{x}$ , on a test may be taken as the sum of the  $p$  observed item scores:

$$\hat{x} = \sum_{j=1}^p x_j, \quad (1)$$

where  $x_j$  is the score on the  $j$ th observed item. Similarly, that with which the writers are most fundamentally concerned, the domain score,  $x$ , is

$$x = \sum_{j=1}^p x_j + \sum_{i=1}^q x_i, \quad (2)$$

the sum of the  $x_{ji}$ , the scores on the  $p$  observed items plus the sum of the  $x_i$ , the scores on the  $q$  remaining hypothetical (unobserved) items in the domain.

To measure confidence in the observed scores as estimates of the domain scores, the correlation  $R$  between  $\hat{x}$  and  $x$  is obtained; clearly, if this correlation is high, the observed score is essentially the desired domain score, whereas if the correlation is low, any inferences about  $x$  and  $\hat{x}$  are to be viewed with suspicion.

Elementary calculation yields the squared correlation  $R^2$  between the observed score (1) and the domain score (2):

$$R^2 = \frac{[p\bar{V}_i + p(p-1)\bar{C}_{ik} + pq\bar{C}_{i.}]^2}{[p\bar{V}_i + p(p-1)\bar{C}_{ik}][p\bar{V}_i + p(p-1)\bar{C}_{ik} + 2pq\bar{C}_{i.} + q\bar{V}_s + q(q-1)\bar{C}_{..}]} \quad (3)$$

where  $\bar{V}_i$  is the mean variance of the observed items,  $\bar{C}_{ik}$  is the mean covariance between the observed items,  $\bar{V}_s$  is the mean variance of the unobserved items,  $\bar{C}_{..}$  is the mean covariance between the observed items, and  $\bar{C}_{i.}$  is the mean (cross) covariance between the observed items and the unobserved items. To clarify these symbols, typical elements of the variance-covariance matrix of all  $p + q$  items in the domain are illustrated in Figure 1.

To simplify the algebra,  $q$  has been kept finite thus far; consistent with the notion that the number of potential items in a domain is indefinitely large, one may take the limit of (3) as  $q \rightarrow \infty$ , and thus obtain:

$$R^2 = \frac{p^2\bar{C}_{i.}^2}{[p\bar{V}_i + p(p-1)\bar{C}_{ik}]\bar{C}_{..}} \quad (4)$$

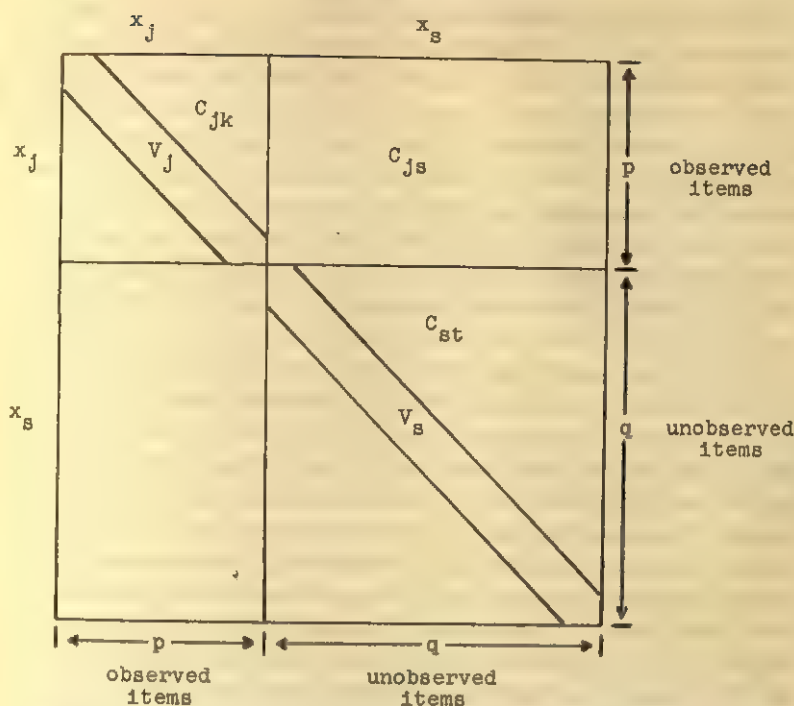


Figure 1. Typical elements in the symmetric variance-covariance matrix of the items in a domain. The subscripts  $j$  and  $k$  refer to observed items; the subscripts  $s$  and  $t$  to hypothetical unobserved items.

It is not possible to evaluate (4) because of the unknown terms  $\bar{C}_{i.}$  and  $\bar{C}_{.i}$ . Thus the crucial point of the development is reached; certain "assumption(s)" about these unobserved quantities must be made. For this purpose, one may take

$$\bar{C}_{i.}^2 = \bar{C}_{ik}\bar{C}_{.i}, \quad (5)$$

When (5) is substituted in (4), it is found that

$$R^2 = \frac{p^2 \bar{C}_{ik}}{p \bar{V}_i + p(p-1) \bar{C}_{ik}}, \quad (6)$$

the basic formula in terms of observables for the squared coefficient of domain validity. This formula may look more familiar if one remembers that its denominator is simply  $V$ , the variance of  $\hat{x}$ , and that  $\bar{C}_{ik} = (V - \sum V_i)/(p(p-1))$ . Hence (6) becomes

$$R^2 = \text{alpha} = \left( \frac{p}{p-1} \right) \left( 1 - \frac{\sum V_i}{\bar{V}} \right), \quad (7)$$

Cronbach's (1951) classic coefficient alpha, the generalized Kuder-Richardson formula 20. The square root of (7),  $R$  itself, is then the correlation between  $\hat{x}$  and  $x$ —Tryon's coefficient of domain validity.

The writers have just engaged in one of the favorite indoor sports of psychometricians: producing a new development for the Kuder-Richardson formula. However, to qualify for serious consideration, new derivations must surely additionally obey the rule of making less restrictive assumptions than previous treatments; as is well known, this classic formula has a history of being developed with unnecessarily restrictive assumptions. Thus, one should look carefully at the crucial assertion of equation (5).

First, let the assumptions that have *not* been made be explicitly stated. Nothing has been said about the individual means, variances, and covariances of the items, or about the internal structure of the items. Additionally, there has been no suggestion that the unobserved items are "parallel" to the observed items; none of the so-called "equivalence" assumptions is set forth.

There is one key statement in (5) concerning the cross-covariance between the observed and the unobserved items: the mean cross-covariance is equal to the geometric mean of the two mean within-covariances. It is suggested that (5) is not really an assumption, but may be viewed as perhaps the simplest possible *definition* serving to link the unobserved items as belonging to the domain inducible from the observed items.

In fact, it is suggested that (5) may be the only reasonable definition in this context, for anything else appears to be unreasonable. If one considers taking the inequality  $\bar{C}_{i.}^2 < \bar{C}_{i.}\bar{C}_{..}$  as a definition, it is clear that the association between observed and unobserved items is generally lower than that within each item set separately—an outcome which surely denies that the hypothetical unobserved items belong (in general) to the domain implied by the observed items.

On the other hand, if one took the inequality  $\bar{C}_{i.}^2 > \bar{C}_{i.}\bar{C}_{..}$ , one would be saying that the association between observed and unobserved items is generally higher than that within each item set separately. In any case, asserting that  $\bar{C}_{i.} > \bar{C}_{i.}\bar{C}_{..}$  is essentially impossible, for it is clear that such a statement would soon render the domain item covariance matrix in Figure 1 indefinite. Indeed, it is easy to calculate the maximum possible  $\bar{C}_{i.}^2$  by letting  $R^2 = 1$  in (4):

$$\text{Max } (\bar{C}_{i.}^2) = \left( \frac{p-1}{p} \right) \bar{C}_{i.}\bar{C}_{..} + \left( \frac{1}{p} \right) \bar{V}_i \bar{C}_{..}, \quad (8)$$



a quantity which rapidly reduces to that of the writers' basic definition (5) as the number of observed items becomes large.

The reader may have the impression that the above treatment constitutes a derivation of the formula for the generalized Kuder-Richardson *reliability* coefficient. This impression is not true, for, until this point, nothing has been said explicitly about reliability of measurement. The development has been entirely in the context of Tryon or Cronbach's profoundly simple and profoundly compelling concepts of domain validity or generalizability. Substantial effort suggests that it is impossible to derive (6) or (7) as a reliability coefficient without bringing to bear the ominous equivalence assumptions that have plagued the classic theory of reliability. Tryon (1957), for once bowing to tradition, first developed  $R^2$  as a reliability coefficient and was forced to invoke the usual restrictive assumptions; he then proceeded to derive his domain validity coefficient, ". . . a statistic that is more meaningful than the reliability coefficient . . .," from his earlier development of the reliability coefficient. In contrast, but following Tryon's algebra closely, the writers have gone directly to his "more meaningful" coefficient without stopping along the way to pick up the excess baggage of the conventional assumptions of classic reliability theory.

#### REFERENCES

- Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, 16, 297-334.
- Cronbach, L. J., Rajaratnam, Nageswari, and Gleser, Goldine C. Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, 1963, 16, 137-163.
- Tryon, R. C. Reliability and behavior domain validity: Reformulation and historical critique. *Psychological Bulletin*, 1957, 54, 229-249.



## AN EMPIRICAL INVESTIGATION COMPARING THE EFFECTIVENESS OF FOUR SCORING STRATEGIES ON THE KUDER OCCUPATIONAL INTEREST SURVEY FORM DD<sup>1</sup>

STEPHEN OLEJNIK

Michigan State University

ANDREW C. PORTER<sup>2</sup>

National Institute of Education

Responses to the Kuder Occupational Interest Survey by 3983 males from nine occupational groups provided the data to compare the accuracy of four scoring strategies: lambda coefficients, currently used by the test publisher, chi-square weights and two applications of multiple discriminant analysis. An analysis of variance test of the percentage of individuals correctly identified by each technique indicated that no statistically significant differences existed between the strategies. The study therefore provided empirical evidence supporting the continued use of the lambda weighting procedure for the scoring of the Kuder OIS.

IMPROVING the discriminatory accuracy of interest surveys has been a problem of concern to measurement theorists for many years. One approach taken to increase the effectiveness of existing instruments has been the development of new scoring techniques (e.g., Findley, 1956; Porter, 1967; Kuder, 1966). As a consequence, a number of promising scoring procedures have been suggested but efforts to identify the optimal method for a given instrument have been lacking.

The purpose of the present investigation was to compare the effectiveness of four strategies for scoring the Kuder Occupational

<sup>1</sup> Paper presented at the 1974 NCME meeting in Chicago, April, 1974.

<sup>2</sup> On leave from Michigan State University.

Interest Survey Form DD. More specifically, techniques which were considered included: Lambda coefficients, the procedure currently used by the publisher; chi-square weights as developed by Porter (1967); and multiple discriminant analysis using occupational scores generated by (a) lambda coefficients, and (b) chi-square weights. Loadman (1972) conducted a similar study which considered a pattern analytic approach rather than discriminant analysis with chi-square occupational scores for the scoring of the Kuder OIS. His comparisons of the four scoring techniques may have been misleading, however, since in his study some of the procedures were cross-validated while others were not. In addition to correcting the problems of cross-validation, the present study also provided for a re-analysis.

### *Format of the Kuder OIS*

The Kuder OIS attempts to identify occupational interests by analyzing responses to one hundred items, each consisting of a set of three activities. Each triad representing an item is presented to the individual in the format shown in Figure 1. From each group of activities the respondent is instructed to select the activity he prefers most and that which he likes least. The two responses per item can be then summarized by one of the following patterns: 1-5, 1-6, 2-4, 2-6, 3-4, or 3-5. Scoring the instrument is problematic since every response is correct if answered sincerely.

### *Scoring Procedures*

#### *Lambda Coefficients*

To solve the scoring problem for the OIS, the publishers (Science Research Associates (SRA) currently use a lambda coefficient as a measure of the similarity of interests between an individual and a particular occupational group. The technique was developed by Kuder (1966) based on the research of Clemans (1958) who had suggested that the relationship between an item and a criterion could be measured by the ratio of a point biserial to the maximum point biserial correlation. In calculating the point biserial correla-

Figure 1.

	Most	Least
Activity 1	(1) 0	(4) 0
Activity 2	(2) 0	(5) 0
Activity 3	(3) 0	(6) 0



tion the dichotomous variable is the selection or nonselection of the 600 possible response patterns, while the continuous variable is the proportion of the criterion group selecting each of the 600 response patterns. Thus the correlation is computed between a vector of 600, 1's and 0's corresponding to the selected response patterns of the individual and a vector of 600 proportions each associated with a separate response pattern. The maximum point biserial correlation is computed based on the selection of the highest response pattern proportion for each item across all 100 items. The ratio of the two correlations produces the lambda coefficient which has an upper limit of 1.00 and is unaffected by the homogeneity of the particular criterion group.

For each occupation a set of lambda weights is calculated. The computational formula used in deriving the lambda weights for a particular item ( $i$ ) and response patterns ( $j$ ) may be presented as:

$$\lambda_{ij} = \frac{P_{ij} - \bar{X}}{\sum_{i=1}^{100} \max_j (P_{ij} - \bar{X})} \quad (1)$$

where  $P_{ij}$  is the sum of the proportion of individuals in a particular criterion group who select the activity most liked plus the proportion of the individuals selecting the activity least liked which make up response pattern ( $j$ ) for item ( $i$ ).  $\bar{X}$  is the average value of the  $P_{ij}$  across all 600 possible response patterns,

$$\bar{X} = \frac{\sum_{i=1}^{100} \sum_{j=1}^6 P_{ij}}{600} = .667.$$

The denominator of equation (1) can be rewritten as  $\sum_{i=1}^{100} \max_j P_{ij} - 100\bar{X}$ . That is the largest  $P_{ij}$  for each of the 100 items are summed together and reduced by a factor of 100(.667) or 66.7. It should be noted then, that the denominator of the equation remains constant for the computation of all 600 lambda weights for a particular criterion group but could vary across criterion groups. The computation of the lambda weight is such that the sum of 100  $\lambda_{ij}$ 's results in the lambda coefficient.

The similarity of an individual's interests with a particular occupational group is estimated by the lambda coefficient. The occupational group in which the individual lambda coefficient is the highest is designated as the most compatible group.

*Chi-Square Weights*

Another approach for developing item response pattern weights for the Kuder OIS was suggested by Porter (1967). Rather than considering one occupational group at a time, as in the lambda procedure, the chi-square technique examines the responses made by several criterion groups. For each of the 100 items on the instrument, a contingency table consisting of a simultaneous breakdown of subjects by occupations and response patterns is constructed. The weights assigned per response pattern per occupation for each table are calculated using the following formula:

$$W_{ij} = \frac{\left(z_{ij} - \frac{X_i Y_j}{\sum X_i}\right)^2}{\frac{X_i Y_j}{\sum X_i}} \times (\text{sign of the unsquared numerator})$$

where  $i = 1 \dots I$  and  $j = 1 \dots J$ ;  $I$  denoting the number of occupations and  $J$  denotes the number of response patterns.  $z_{ij}$  is the number of responses made by the  $i$ -th group to the  $j$ -th response pattern,  $X_i$  is the total number of subjects in the  $i$ -th group,  $Y_j$  is the total number of individuals selecting the  $j$ -th response pattern and  $\sum X_i$  is the total number of subjects in the sample.

Thus for each of the occupations considered, a fractional weight is calculated for each of the possible response patterns. An individual's score for an occupation is the sum of the chi-square weights associated with the individual's responses to the 100 items. The higher the total score, the greater the similarity in interests between the individual and a criterion group.

*Multiple Discriminant Analysis*

Still another approach to the scoring problem, and one which has been used with considerable success in other classification problems, is the application of the linear discriminant function. In the present study, this technique was applied using two different sets of variables. The first set of variables was the occupational scores obtained following the lambda weighting procedure and the second set of variables was the occupational scores developed from the chi-square weighting procedure. Following the identification of the best linear combinations of these variables, classification of an individual was obtained by the simple  $d^2$  function, i.e., for each occupation the sum across functions of the squared deviations of an individual's composite score from the mean composite score. The respondents

were classified as belonging to the occupational group corresponding to the smallest  $d^2$  value.

### *Sample*

To develop and test the effectiveness of the four scoring procedures described, the item responses of 3893 males, from nine unequally sized occupational groups, to the Kuder OIS were analyzed. These responses were originally collected by Kuder while developing the instrument and later obtained by Porter for the development of the chi-square scoring procedure. This data was also used by Loadman in his analysis. Although the data are probably too old for use in developing scoring keys for present day use, they did serve as an adequate base for comparison of the four scoring procedures.

The occupational groups selected for study included: pediatricians, veterinarians, physical therapists, x-ray technicians, optometrists, clinical psychologists, social workers, foresters, and auto mechanics. The first five groups were designated as set I and were considered as similar occupations, while the last four groups plus optometrists were considered as dissimilar and labeled set II. Since optometrists appeared in both sets of data, the two sets were not independent. Each occupational group was randomly divided into two halves, A and B; thus two independent groups of data were available for each set. To obtain an estimate of the "true" effectiveness of each scoring procedure, a double cross-validation technique as suggested by Mosier (1951) was followed.

### *Analysis*

To compare the results of the four Kuder OIS scoring procedures, an analysis of variance for mixed models was utilized. The dependent variable in the study was the average (across halves A and B) percentage of correctly identified individuals for an occupation. The problem of nonindependence between sets was solved by using the cross-validated results of half A for optometrists in set I and the cross-validated results of half B in set II. The fixed independent variables were: sets, similar or dissimilar occupations, S; measures, lambda coefficients or chi-square weights, M; and discriminant analysis or not, D. All three fixed independent variables were completely crossed with each other. Occupation was treated as a random independent variable which was nested within S, with five levels per nest, but crossed with the two scoring procedures D and M.

Although occupations were not actually selected at random from a larger pool of occupations, it was felt that by using the Cornfield-Tukey bridge argument (Cornfield and Tukey, 1956) the results of this study could be generalized to similar occupations. Had occupations been treated as fixed, greater power would have resulted in the analysis since the individual test respondent would have been the unit of analysis rather than occupations. The results of such a test, however, would have had very little practical value.

The hypotheses tested included differences in the discriminatory accuracy: between sets of occupational groups, between measures, chi-square and lambda techniques, and between using multiple discriminant analysis or not. Interaction effects between measures and discriminant analysis as well as interaction effects with sets were also tested. Each hypothesis was tested for statistical significance at  $\alpha = .05$ .

### Results

The average percentage of individuals in half A and B correctly classified into their actual occupational groups based on the cross-validated scoring keys are presented in Table 1. The cross-validated average percent correct classifications in the present study were compared with Loadman's (1972, p. 114) quasi-cross-validated results. As was expected the cross-validated results were in each case smaller. For the lambda procedure the shrinkage due to cross-

TABLE 1

*The Average Percentage of Individuals Correctly Classified Into Their Actual Occupational Group from Half A and Half B for Set I and Set II*

		Discriminant Analysis		Non-Discriminant Analysis	
		Chi-Square Weights	Lambda Weights	Chi-Square Weights	Lambda Weights
Set I	Optometrist	62.54	54.19	60.10	63.05
	X-Ray Technician	52.94	30.61	43.08	54.37
	Pediatrician	57.47	44.14	65.28	67.58
	Physical Therapist	42.91	42.19	38.06	49.85
	Veterinarian	78.56	71.98	88.40	60.11
Set II	Clinical Psychologist	71.40	57.54	86.80	71.20
	Auto Mechanic	83.34	83.00	96.00	81.00
	Forester	76.30	68.79	43.56	78.04
	Optometrist	78.00	51.50	38.00	75.00
	Social Worker	53.21	51.63	53.88	70.06



TABLE 2

*ANOVA Table for Mixed Models, Analyzing the Data from Table 1*

Sources	Degrees of Freedom	Means Squares	F	p
S	1	1449.86	2.51	.15
O:S	8	578.07		
D	1	126.59	1.47	.26
M	1	48.44	.48	.52
SXD	1	28.06	.33	.59
SXM	1	85.73	.85	.39
DXM	1	625.84	3.55	.10
SXDXM	1	76.95	.44	.53
OXD:S	8	85.97		
OXM:S	8	100.90		
OXDXM:S	8	176.51		

validation was 5.63 in set I and 1.27 in set II. For the discriminant analysis procedure based on lambda scores the shrinkage was 11.26 in set I and 14.70 in set II. Thus a great deal of bias had entered Loadman's study when he compared his non-cross-validated lambda based results with his cross-validated chi-square based results.

The ANOVA table associated with the data in Table 1 is presented in Table 2. The unweighted average percentage of individuals correctly classified in sets I and II were 56.37 and 68.41 respectively. Although the difference in percentage correct identifications between sets was in the predicted direction, the null hypothesis of no difference between sets was not rejected,  $p < .15$ . A comparison of the average correct classification rates (across the four scoring techniques) among optometrists in set I and set II provided a further test of the hypothesis. Among the homogeneous occupations, optometrists were correctly identified an average of 60.04% of the time, while among heterogeneous occupations, 61.87% of the individuals were correctly classified. This comparison suggested that for the techniques investigated, discrimination among similar occupations was, practically speaking, as accurate as among dissimilar groups.

The unweighted average percentage of individuals correctly identified when multiple discriminant analysis procedures were used was 60.62, while non-use of this technique produced 66.66% correct classifications. This difference was not statistically significant,  $p < .26$ . A comparison of the measures, lambda vs. chi-square showed that for the former an average of 61.29% of the individuals were correctly classified while for the latter an average of 63.49%

correct classifications were made. The null hypothesis of no difference between measures was not rejected,  $p < .52$ .

An average correct classification rate of 65.67% was made with the discriminant analysis based on chi-square scores; 55.56% correct classification rate was obtained with the discriminant analysis based on lambda occupational scores; 61.32% correct classification rate was obtained based on the chi-square scoring techniques used alone and 67.03% correct identification rate was obtained when the lambda coefficient was used alone. Again the analysis did not indicate a statistically significant interaction,  $p < .10$ . Finally no interaction effects with sets, S, were identified.

A further consideration in deciding which scoring technique was the most effective was consistency in the accuracy with which the procedures correctly classified individuals across occupations. A technique which correctly identifies individuals in one or two occupations at a very high rate but classifies individuals in other occupational groups at low rates may not, in the long run, be as valuable as a procedure which consistently classifies individuals at a moderately high rate over all occupations. The hypothesis of equal consistency of accuracy across occupations was tested by a two-way analysis of variance (occupations by scoring procedure). The dependent variable was the absolute value of the difference between the average percentage of individuals correctly classified and the mean average percentage correct classification rate within each scoring procedure (Levene, 1960). The standard deviation of the average percentage of individuals correctly classified as well as the averages of the absolute values of the deviations from the means within each scoring procedure are presented in Table 3. The null hypothesis of equal consistency of accuracy among the four scoring procedures was not rejected for set I. Using the same test with

TABLE 3

*The Standard Deviation and Absolute Error Difference for Each Strategy Used in Levene's Test for Equality of Variance Across the Four Measures*

	Set I		Chi-Square	Lambda
	Discriminant with Chi-Square	Discriminant with Lambda		
Standard Deviation	13.16	15.51	17.87	7.61
Average $\bar{z}$	9.33	11.57	14.73	5.51
	Set II		Chi-Square	Lambda
	Discriminant with Chi-Square	Discriminant with Lambda		
Standard Deviation	11.57	13.26	23.40	4.58
Average $\bar{z}$	8.12	10.75	22.20	3.57

set II, however, the null hypothesis was rejected. The Scheffé post hoc test indicated that the chi-square procedure was less consistent than either the lambda technique or the discriminant procedure based on chi-square occupational scores.

### Summary

The results of the present study as presented in Tables 1 and 2 indicate that no one scoring procedure was superior to the others across all of the occupational groups considered. The lambda weighting procedure, however, correctly identified individuals as belonging to their actual occupational group at the highest rate in six of ten occupations, the chi-square technique in three of ten, the discriminant analysis based on chi-square occupational scores in one of ten. The discriminant analysis technique based on the lambda occupational scores did not have the highest rate of correct classification in any of the occupations studied. In addition, the four scoring procedures did not differ on the criterion of consistency of accuracy for similar occupations, but the lambda and the discriminant analysis based on the chi-square scores procedure were more consistent than the chi-square procedure for the dissimilar occupations.

Considering that the discriminant analysis procedures are more difficult to calculate and have no greater accuracy than the lambda procedure, and since the latter had greater consistency than the chi-square technique, the lambda procedure seems preferable. Furthermore, the lambda coefficient is also preferable to the chi-square procedure since lambda weights are not a function of the occupations being compared while the chi-square weights for an occupation depend on the other occupations in the set being compared.

### REFERENCES

- Clemans, W. V. An index of new criterion relationships. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1958, 18, 1.
- Cornfield, J. and Tukey, J. Average values of means squares in factorials. *Annals of Mathematical Statistics*, 1956, 27, 907-949.
- Findley, W. A rationale for evaluation of item discrimination. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1956, 16, 175-180.
- Kuder, G. F. A comparative study of some methods of developing occupational keys. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1957, 17, 105-114.
- Kuder, G. F. *Kuder Occupational Interest Survey Form DD. General Manual*. Chicago: SRA. 1966.
- Levene, H. Robust tests for equality of variance. In I. Olkin (Ed.),

*Contributions to probability and statistics.* Stanford, California, 1960.

Loadman, W. E. A comparison of several methods of scoring the Kuder Occupational Interest Survey. (Doctoral dissertation, Michigan State University) Ann Arbor, Michigan: University Microfilms, 1972, No. 72-16470.

Mosier, C. I. Problems and design of cross-validation. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1951, 11, 5-11.

Porter, A. C. *A chi square approach to discrimination among occupations using an interest inventory.* Technical report No. 24. University of Wisconsin Center for Cognitive Learning, Madison, Wisconsin, 1967.



## THE 27 PERCENT RULE REVISITED

RALPH B. D'AGOSTINO

Boston University

EDWARD E. CURETON

University of Tennessee

The problem of selecting the size of the tails from a sample drawn from a distribution of test scores for item analysis study is discussed. In the past the selected tails were viewed as independent samples. However, this is not the case. The tails are dependent. Given this we find that each tail should contain about 21% of the sample and not the traditional 27%. Use of 27%, however, is not far from optimal.

In the traditional item analysis situation a test is administered to  $N$  subjects, the resulting scores are arranged in order of magnitude and the tests of these subjects whose scores fell in either the upper or lower tails of the distribution of test scores are further studied item by item. A question that arises here is how large should the tails be—that is, what is the appropriate  $q$  ( $0 < q \leq .50$ ) such that the upper and lower tails to be selected for further study each contain  $q$  of the cases?

An argument for determining  $q$  runs as follows. Assume the underlying distribution of test scores is normal and the sample size is large. Let  $\bar{X}_1$  and  $\bar{X}_2$  be the means from the upper and lower tails of the sample respectively. Then  $q$  should be determined so that the critical ratio

$$CR = \frac{\bar{X}_1 - \bar{X}_2}{SE(\bar{X}_1 - \bar{X}_2)} \quad (1)$$

is maximized. Here  $SE(\bar{X}_1 - \bar{X}_2)$  is the standard error of  $\bar{X}_1 - \bar{X}_2$ . Now  $CR$  of (1) will vary from sample to sample due to sampling

fluctuations. So it is more appropriate to first take expectations and then maximize. This means for large samples the maximizing of

$$E(CR) = \frac{2f\sigma/q}{SE(\bar{X}_1 - \bar{X}_2)} \quad (2)$$

where  $f$  is the unit normal ordinate at the upper standardized baseline point  $z$  which separates the tail from the rest of the distribution and  $\sigma$  is the standard deviation of the underlying distribution of test scores. That is,  $z$  and  $f$  are defined by

$$q = \int_z^{\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx \quad (3)$$

and

$$f = \frac{e^{-z^2/2}}{\sqrt{2\pi}}. \quad (4)$$

In previous works (Cureton [1957] and Kelley [1939]) it was implicitly assumed that the upper and lower tails constitute independent subgroups. However, this is not the case. *The upper tail observations are correlated with the lower tail observations even for large samples* (Mosteller, 1946). Because of this the standard error of  $\bar{X}_1 - \bar{X}_2$  displayed in the above references is incorrect and the  $q$  which maximizes the critical ratio is not .27 as was previously believed. As we show below it is about .21.

Using the method of Chernoff, Gastwirth and Johns (1967) the correct standard error can be shown to equal, asymptotically (i.e., for large samples),

$$SE(\bar{X}_1 - \bar{X}_2) = q\sqrt{\frac{\sigma}{N}} \sqrt{A} \quad (5)$$

where

$$A = 2q + 2zf + z^2(1 - 2q) - (2f + z(1 - 2q))^2. \quad (6)$$

The term to maximize is thus

$$E(CR) = \frac{2f}{\sqrt{A}} \sqrt{N} \quad (7)$$

or equivalently the term to maximize is

$$\frac{4f^2}{A}. \quad (8)$$

This last term was computed for  $q = .01(.01).50$ . It was found to

attain its maximum around  $q = .21$  and  $.22$ . In particular some entries are

<u><math>q</math></u>	<u><math>4f^2/A</math></u>
.19	1.9293
.20	1.9325
.21	1.9338
.22	1.9336
.23	1.9320
.27	1.9147
.50	1.7519

Using four-point interpolation (8) attains its maximum for  $q$  to three decimal places at  $q = .215$  and here  $4f^2/A = 1.9339$ . The optimal  $q$  is actually slightly closer (i.e., consideration of fourth decimal place) to  $.21$  than  $.22$ . Thus instead of the 27% rule, more appropriately it is the 21% rule. However, as is clear from the above entries little is lost by using  $q = .27$ .

The arguments of Cureton (1957) and Kelley (1939) would be correct if instead of selecting the tails of a sample, two independent samples were drawn from the tails of the complete distribution of test scores. This, of course, is usually impossible for the actual distribution is unknown. The best that is possible would be to select independent samples from the tails of some concomitant variable where we can view the test scores as a dependent measure. This approach is exactly the extreme group approach discussed by Feldt (1961). In this case the optimal tail size is around  $.27$  if the correlation between the concomitant variable and the test scores is small (i.e., around  $.10$ ). Under normality this is implying independence. As the correlation increases the optimal tail size decreases. From above it appears to follow that if the concomitant variable and the test scores have correlation one then the optimal tail size is round  $.215$ .

We should also mention that the results of Ross and Weitzman (1964) do not disagree with our results. For they show that the optimal size of the tails (under bivariate normality) for estimating the tetrachoric correlation is  $.27$  when again the variables are in fact uncorrelated. The optimal tail sizes differ when the correlation is not zero.

Finally, it should be stressed that the above demonstration raises a serious question regarding the techniques presently employed for item discrimination analysis based on answers to an item in high

and low criterion groups. The techniques are the Chi-square test in a four-fold table or the two sample  $t$  test. The present tests assume two independent groups. As indicated above this is not so. Either the validity of the tests need to be demonstrated or new techniques taking into account the dependency need to be produced. The present authors are working towards these.

### REFERENCES

- Chernoff, H., Gastwirth, J. L. and Johns, M. V. Asymptotic distribution of linear combinations of order statistics. *The Annals of Mathematical Statistics*, 1967, 31, 52-72.
- Cureton, E. E. The upper and lower twenty-seven percent rule. *Psychometrika*, 1957, 22, 293-296.
- Feldt, L. S. The use of extreme groups to test for the presence of a relationship. *Psychometrika*, 1961, 26, 307-316.
- Kelley, T. L. The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, 1939, 30, 17-24.
- Mosteller, F. On some useful "inefficient" statistics. *The Annals of Mathematical Statistics*, 1946, 17, 377-408.
- Ross, J. and Weitzman, R. A. The twenty-seven percent rule. *The Annals of Mathematical Statistics*, 1964, 35, 214-221.



## A MODEL FOR PSYCHOMETRICALLY DISTINGUISHING APTITUDE FROM ABILITY

SUSAN E. WHITELY AND RENÉ V. DAWIS<sup>1</sup>

University of Minnesota

It is widely agreed that current ability measures reflect a complex interaction of environment with genetic potential. This leads to a basic measurement problem since persons with the same measured ability may vary widely in potential due to nonequivalent learning opportunities. The purpose of this paper is to present a model which may hold some promise in psychometrically distinguishing ability (current status) from aptitude (potential). Data on spatial reversal performance are analyzed according to the model to illustrate how some of the practical problems may be solved.

It is widely agreed that current ability measures reflect a complex interaction of environmental and genetic factors. The literature on general intelligence, for instance, has unequivocally demonstrated that test performance is highly influenced by membership in a culture or subculture such as a race or socio-economic class. Although there still is much controversy over whether these subcultures differ genetically (Herrnstein, 1971; Jensen, 1969), it is known that exposure to more advantageous environments can increase IQ (cf. Lee, 1951). Thus, with general intelligence (and probably many other abilities) the particular learning experiences and opportunities an individual encounters has a significant in-

---

<sup>1</sup> This research was supported by contract number N00014-68-A-0141-0003 from the Office of Naval Research to the Center for the Study of Organizational Performance and Human Effectiveness, Department of Psychology, University of Minnesota. Reprints may be obtained from: René V. Dawis, Department of Psychology, University of Minnesota, Minneapolis, Minnesota 55455.

fluence on his measured ability. This leads to a basic measurement paradox since individuals who show the same measured ability may have different potential if their learning experiences have varied widely. The problem is to find a method of determining which individuals have undeveloped potential so that they may be provided more effective educational experiences or subjected to more appropriate selection criteria.

The major purpose of this paper is to present a method which psychometrically distinguishes between ability (current status on a test) and aptitude (potential). Tests measuring cognitive ability factors have not, as yet, been developed to the point where the general utility of this approach can be assessed. However, data on a psychomotor ability will be presented to illustrate the feasibility of the approach and to suggest ways to solve some of the more practical problems in the application of the technique.

### *A Conceptualization of the Relationship of Ability to Aptitude*

At the conceptual level, it is suggested that ability can be defined as current status and aptitude can be defined as potential status under conditions optimally favorable to the development of the ability. Direct assessment of aptitude, then, would imply equally favorable learning experiences for all individuals. Obviously, this is not the case, but it is theoretically possible to distinguish between individuals having the same ability but different aptitudes by directly measuring the modifiability of ability. That is, when two individuals within the same current level of ability are given equivalent intervention (e.g., practice), the individual with the greater aptitude should show a faster rate of change in ability than the individual with lesser aptitude.

It is important to consider this general conceptualization of aptitude measurement in the context of previous research, since many investigators have found the measurement of change to be not only statistically complex, but of very questionable utility. Woodrow's (1939) finding that improvement over practice is not the same as learning ability has been largely unrefuted by subsequent investigations (see Cronbach and Snow, 1969, for summary). Woodrow found gain to be both task-specific and not correlated with a general ability factor. Since gain scores do not seem to reflect learning ability, it would appear that measuring modifiability would not be a useful technique in discriminating aptitude from ability.

However, this conclusion may be premature because correlation

with gain scores apparently depends on two important variables. First, the stage of practice over which gain is computed may moderate the relationship between gain score and other variables. Both Woodrow (1939) and more recent studies (e.g., Dunham, Guilford and Hoepfner, 1968) have found that the factorial composition of a task may change over stages of practice. Gain at different stages of practice, then, may be expected to depend on different abilities. It is interesting to interpret the Woodrow (1939) study from this perspective. Woodrow used 66 trials for four practiced tests and found the final scores to be less correlated with cognitive factors than the initial scores. With long practice periods, subjects approach the mastery level and, as Jones (1970a) has noted, the variation among subjects at this level becomes increasingly due to error (unreliability). Both unreliability and motor ability may have accounted for a large share of the final task variation among Woodrow's subjects. More importantly, gain computed over such a long interval of practice could be expected to reflect task-specific mastery components rather than a more general learning ability.

A second important moderator of correlation with gain scores is the relationship of gain to initial status. Gain correlates negatively with initial status in most learning tasks, so that those with the least efficient initial performance gain the most. Variability in gain scores, then, reflects two confounding factors, stage of practice and initial status.

A successful distinction between aptitude and ability will have to account for these two important variables. Stages of practice should be carefully studied so that gain can be measured during those stages most likely to reflect learning ability. Also, the confounding effect of initial status should be controlled. The model to be proposed here directly controls the influence of initial status and uses a special technique, molar correlation analysis (Jones, 1962; 1970a; 1970b), to investigate stages of practice.

It is hypothesized that aptitude can be reflected by a linear combination of two measures, ability (current status) and modifiability. The simplest way of expressing this relationship is by the equation for a straight line,  $y = a + bx$ . The symbols in the equation are defined as follows: (1) the constant,  $a$ , is the initial status on the ability test; (2)  $bx$  is the modifiability of the ability test scores; and (3)  $y$  is the aptitude when measured ability is at asymptotic value. Modifiability has two separate components. One of these,  $x$ , refers to either the graded quality or amount of intervention between ability estimates, while  $b$  refers to the rate of ability change

observed. An individual's aptitude, then, is conceived of as a fluid quantity, characterized by both his initial status,  $a$ , and sensitivity to intervention,  $b$ .

It follows, then, that in order to have a measurement which reflects aptitude, it is necessary to have at least two measurements of ability, one before and one after a standardized intervention (fixed value of  $x$ ). For prediction, ability and modifiability would be two different independent variables in a multiple regression equation, weighted according to their relative importance to the criterion to be predicted.

This simple formulation, with the hypothesized role of gain as a predictor, has interesting implications. Cronbach and Furby (1970) have suggested that "residualized gain scores" rather than raw gain should be used if the goal is to identify individuals with undeveloped potential. Thus, only "unexpected" gain would be used to select such individuals. When raw gain is used in a regression equation with initial status, it can be shown that the beta weight for raw gain is linearly related to the correlation between residualized gain and the criterion.<sup>2</sup>

Similarly, the use of a regression equation also implies that raw gain need not be correlated with a criterion measure to yield increased predictability. Gain may function as a *suppressor variable* through its correlation with initial status. Theoretically then, gain as a modifiability measure can be used to correct initial status scores to provide a better reflection of aptitude.

---

<sup>2</sup> Consider the following regression equation:

$$x_1 = \beta_2 x_2 + \beta_3 x_3,$$

where 1 refers to the criterion deviation score, 2 refers to the deviation on initial status and 3 refers to the deviation of the raw gain score. The beta weight for raw gain can be computed as follows:

$$\beta_3 = \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2},$$

and the part correlation of the criterion and raw gain, removing the effects of initial status (definitionally the correlation of the residualized gain score and the criterion) is as follows:

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1 - r_{23}^2}}.$$

Then

$$r_{13.2} = \beta_3 \sqrt{1 - r_{23}^2}.$$



There are several problems which directly touch on the feasibility of utilizing such an approach to the measurement of aptitude. The first, and most obvious, concerns the extent to which ability scores can show gains from very short intervention periods. Previous studies on coaching (see Anastasi, 1958, for summary) indicate that large gains can be made, and that there are individual differences with respect to gains. Apparently, students from the most deficient environments show the largest gains. It is not clear, however, whether this is due to the correlation of gain with initial status or if there are also larger unexpected gains for such a group.

A second problem concerns the degree to which modifiability of the test score parallels the latent ability trait. Basically this question concerns the relationship between the asymptotic value obtained in the prediction and the latent aptitude. Practically speaking, in the long run, the degree of correspondence here will be determined by the extent to which modifiability scores lead to increased predictability of achievement. However, in the short run, there is a design problem with respect to the degree and nature of the standardized intervention. For instance, little correspondence between latent aptitude and asymptotic test score would be expected when the intervention utilizes the same items that are used for final testing. The asymptotic test score would then depend more on rote memory than on aptitude. Similarly, measuring gain over long intervention periods would be expected to be less correspondent to latent trait modifiability and more specific to the predictor.

A set of related problems concerns the kind of rate measure to use to provide increased reflection of aptitude. The most critical of these problems concerns the relationship of rate measurements to the true shape of the individual ability curves. Most likely, this curve is *S*-shaped such that slope between any two points varies over the course of intervention. If initial status is near the bottom of this curve (large undeveloped potential), the instantaneous rate (derivative) will start out at a low level and then increase to a maximum rate, followed by a decrease in rate until asymptotic value is reached. Thus, it is not necessarily the case that for individuals at the same initial status, the one with the highest aptitude will have the highest modifiability. It depends on what point of the curve is being observed.

Figure 1 presents ability curves and observed rates of change for two hypothetical individuals. Two individuals may have the same observed rate of change (observed between two distance points over intervention) when one has a decreasing instantaneous rate

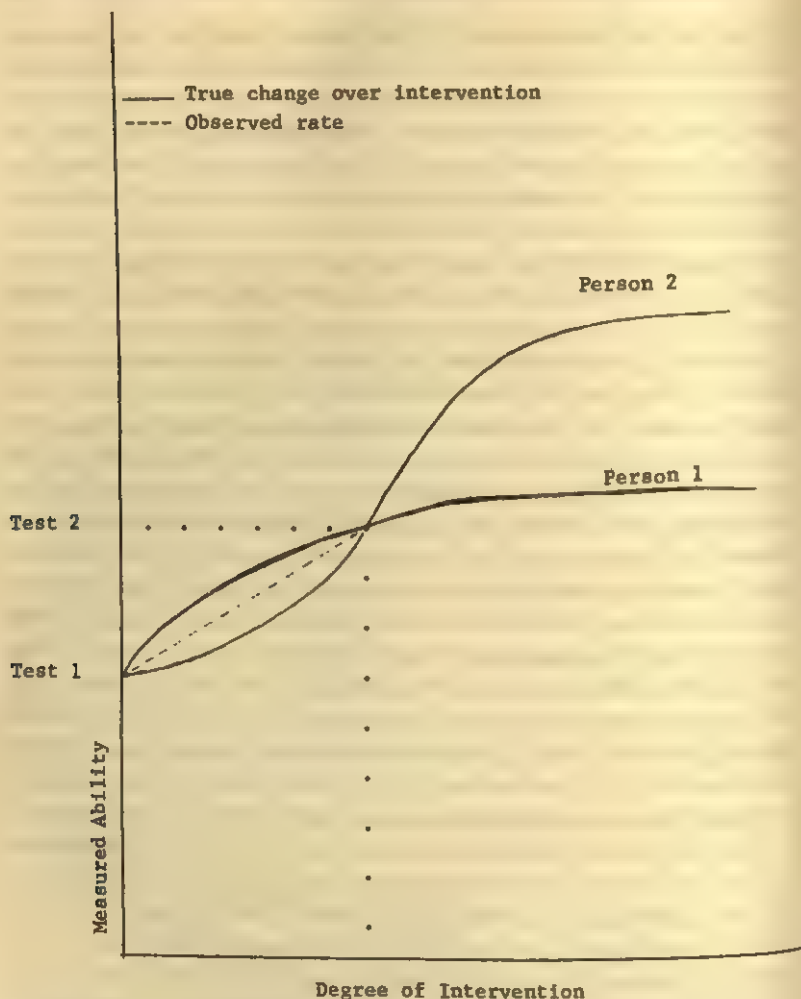


Figure 1. Hypothetical ability curves and observed rates of change for two individuals.

and the other an increasing rate. The one with the increasing rate (Person 2) will reach the higher asymptotic level, but this will not be detectable by gain between these two points.

There are at least two possible approaches to this problem. One is to take several slope (rate) measurements over increasingly better interventions. This may be impractical because it is time-consuming for complex abilities and may pose insurmountable difficulties with respect to precision of measurement of ability scores. A second

approach is to use only one segment of intervention, but to select this intervention segment such that the reflection of aptitude by modifiability is maximized in the measurements. In the section that follows, molar correlation analysis (Jones, 1962; 1970a) is used on psychomotor ability data to demonstrate how this selection may be maximized in a population of individuals.

A final problem which may affect the utility of the proposed model concerns how rate is to be computed. If a single intervention session is used, i.e., observing ability only twice, there is no obvious unit to use on the abscissa. When rate is to be used in a regression equation with initial status, there are three possible ways of computing it: (1) slope, (2) score ratio and (3) gain score. To compute the first rate index, slope, some measurement of performance must be taken during intervention. It may be feasible to generate such measures, such as time spent in practice, but the interpretability is not always clear. The second index, score ratio, can be used on tests that provide ratio scale measurement. Although no currently existing ability test provides ratio scale measurement, there are new scaling methods (e.g., Wright and Panchapakesan, 1969) which can potentially allow the computation of score ratios. The third possible index to use is raw gain score. Since the gain score is to be used in a regression equation, as proposed above, many of the objections to raw gain score are eliminated.

However, no matter which index of rate is used, the reliability of these measurements from equivalent forms should be investigated during the development of the tests. To depend on tests developed according to classical criteria of equivalent forms leads to paradoxes in the estimation of the reliability of rate scores. That is, gain is not independent of measurement error.

Information on the general utility of this approach on complex cognitive abilities apparently must await further developments. However, data on a psychomotor ability are presented below to illustrate the relationship of modifiability to prediction and to suggest internal criteria for the selection of an appropriate segment of intervention.

### *The Predictability of Spatial Reversal Performance*

#### *Materials*

A task which requires spatial reversal ability, tracing a simple figure in a mirror-blind apparatus, was used to provide ability status and modifiability measurements. In the mirror-blind ap-

paratus, the only visual cues are completely reversed from normal eye-hand coordination tasks. This task has been shown to be highly influenced by experience, although individual differences do persist (P. W. Fox, personal communication).

The figure to be traced for the predictor (status and modifiability) measurements was a "zig-zag" line that required reversals in only two different directions. Both reversals were at 45 degree angles. The criterion task to be predicted by these status and modifiability measurements was the tracing of a more complex figure, a six-pointed star, in the mirror-blind apparatus. The star was constructed such that the role of spatial reversal ability would be maximized and task overlap between the zig-zag line and the star with respect to specific reversals would be minimized. On the star, no two reversals were in the same direction at the same angle. Also, none of the reversals on the star was in the same direction as on the zig-zag line. To equate the role of motor speed on these tasks, the star and zig-zag line were equated for total number of reversals and distance between reversals. The resulting correlations between the predictor measurements on the zig-zag line and criterion measurements on the star should then be due to motor reversal ability. The general question asked of the data, then, is: does modifiability on a specific measure of an ability (similar to coaching on a test with homogeneous items) add anything to the prediction of a complex task assumed to load heavily on the ability? If so, then modifiability on the zig-zag line should add to predictability on the star, in the mirror-blind task.

### *Subjects and Procedure*

The subjects were 49 college sophomores enrolled in elementary psychology courses at the University of Minnesota. Four subjects were dropped from the experiment: two because of equipment failure, one for exceeding the five minute time limit, and one for taking a drug known to influence psychomotor performance.

Each subject was given 10 successive trials on tracing the zig-zag line in the mirror-blind apparatus. Immediately following these trials, the star was traced for one trial in the mirror-blind apparatus. Time, in seconds, was recorded for each trial. High scores, on both predictor and criterion, indicate inefficient performance.

### *Results*

Table 1 presents the means, standard deviations, and correlations between trials of the spatial reversal task on the zig-zag line. It



TABLE I  
Means, Standard Deviations and Intertrial Correlations

Trial	1	2	3	4	5	Correlations				8	9	10	Criterion	$\bar{X}$	SD
						6	7	8	9						
1	—	.81	.66	.36	.38	.34	.34	.19	.23	.25	.23	.25	.23	97	48
2	—	—	.72	.44	.36	.38	.41	.24	.34	.35	.28	.35	.28	61	27
3	—	—	—	.60	.61	.67	.68	.53	.59	.54	.37	.48	.37	48	18
4	—	—	—	—	.55	.50	.58	.67	.67	.62	.52	.47	.52	47	20
5	—	—	—	—	—	.62	.69	.70	.65	.68	.27	.39	.27	39	14
6	—	—	—	—	—	—	.77	.66	.64	.65	.36	.35	.36	35	10
7	—	—	—	—	—	—	—	.70	.66	.66	.37	.32	.37	32	10
8	—	—	—	—	—	—	—	—	.82	.76	.25	.31	.25	31	10
9	—	—	—	—	—	—	—	—	—	.81	.41	.29	.41	29	8
10	—	—	—	—	—	—	—	—	—	—	.44	.26	.44	26	6

can be seen that both the mean number of seconds to complete the task and the variability decrease over trials. The correlations between status on the predictor trials and the criterion are also presented on Table 1. All correlations were significant ( $p < .05$ ) except for Trial 1 ( $p = .06$ ). The highest correlation for these status measurements was at Trial 4 (.52).

An inspection of the intertrial correlations presented on Table 1 shows that the correlations display superdiagonal form (Jones, 1962). That is, as one moves down the columns of the correlations matrix or across rows to the left, the correlations increase in size. Adjacent measurements of reversal performance, then, correlate more highly than remote ones. Jones (1970b) has found this pattern to be the general rule over trials of practice, with the exception of very simple psychomotor tasks.

Table 2 presents a decomposition of the total correlation matrix of the ten predictor trials into rate and terminal process components, as suggested by Jones (1970a). Jones hypothesizes that for intertrial correlation matrices having superdiagonal form, the consistency of performance over trials is due to some combination of a rate and a terminal process. The terminal process is defined as the relative ordering of subjects when all have reached their terminal positions. The extent to which the rate process exists between trials indicates the extent to which individual differences in rate of change are contributing to the consistency of performance. The rate process is defined as the residual correlation between trials after partialing out the correlations due to terminal position. Jones (1970a) suggests that the rate process is usually strongest during the early stages of practice and gradually decreases as the terminal

TABLE 2

*Decomposition of the Intertrial Correlations into Rate and Terminal Process Components*

Trial	1	2	3	4	5	6	7	8	9
1		.09	.14	.16	.17	.16	.17	.19	.20
2	.72		.19	.22	.24	.23	.23	.27	.28
3	.52	.50		.33	.37	.35	.36	.41	.44
4	.20	.22	.27		.42	.40	.41	.47	.50
5	.21	.12	.24	.13		.44	.45	.52	.55
6	.18	.15	.32	.10	.16		.43	.49	.53
7	.19	.18	.32	.17	.24	.34		.50	.53
8	.00	.01	.12	.13	.18	.17	.20		.62
9	.03	.06	.15	.17	.10	.11	.13	.20	

Note.—The terminal process appears above the main diagonal; the rate process below it.

process takes over in later stages of practice. Since true asymptote is never reached, Jones suggests that the last trial in the matrix be used to estimate terminal position.

In Table 2, the part of the intertrial correlation due to terminal position (Trial 10), appears above the diagonal, while the rate process appears below it. It can be seen that the terminal process is stronger late in practice, by the increase of correlations moving down the columns and across the rows. The rate process, on the other hand, is strong before Trial 4, but then becomes small and irregular after this point. Thus, the patterns of performance indicate that there are consistent differences between subjects before Trial 4, which are independent of their terminal positions. This indicates that subjects are changing at different rates. The terminal process starts at Trial 4, but stays at a constant strength until Trial 7. At Trial 7, the terminal process again begins to increase.

In Table 2, the triangles enclose what appear to be different stages of practice for the series of trials. Moving down the main diagonal, the correlations in the first triangle on either side mark the termination of the rate process. The second set of triangles designate intermediate trials in which consistency is mainly due to terminal process, but the terminal process is not increasing with practice. This is a somewhat unusual stage of practice since terminal process usually increases regularly. The third set of triangles marks late stages of practice in which the terminal process increasingly determines consistency between trials.

The influence of stages of practice in determining the function of gain in predicting a criterion was investigated by computing the patterns of predictability from gain and initial status over the total trial matrix. With the exception of Trial 1, which did not correlate with the criterion, and Trial 10, the last trial, each trial was treated as an initial status measure. Gain was computed separately between each initial status trial and each succeeding trial of practice.

Table 3 presents the increased percentage of variance accounted for in the criterion when gain score is added to initial status in a regression equation. It can be seen that the highest increment in prediction occurs when Trial 2 is initial status and gain is computed between Trial 2 and Trial 4. The multiple  $R$ , not reported in Table 3, was .52. Also significantly increased prediction at the .01 level is the gain between Trial 3 and Trial 4, using Trial 3 as initial status. Change between Trial 2 and Trial 3, both within the rate process, added nothing to predictability.

TABLE 3

*Change in Percentage of Variance Accounted for by the Addition of Gain Scores in the Regression Equations*

Gain from Trial	2	3	4	5	Gain to Trial		8	9	10
					6	7			
2	—	.06	.19**	.03	.07	.08	.03	.11*	.13*
3	—	—	.14**	.00	.02	.03	.00	.05	.08*
4	—	—	—	.00	.01	.01	.01	.01	.02
5	—	—	—	—	.06	.07	.01	.09*	.12*
6	—	—	—	—	—	.02	.00	.05	.07
7	—	—	—	—	—	—	.00	.05	.07
8	—	—	—	—	—	—	—	.13*	.15**
9	—	—	—	—	—	—	—	—	.19**

\*  $p < .05$ .

\*\*  $p < .01$ .

Change occurring late in practice, from Trial 8 and beyond, also yields significant beta weights for gain scores. It can be seen from Table 3 that all gain computed beyond Trial 8 yielded significantly increased predictability. It is interesting to note, however, that gain from the rate process trials, Trial 2 and Trial 3, to any of the intermediate trials, Trial 5, Trial 6, Trial 7, and Trial 8, did not produce any increased predictability. However, when gain is computed to Trial 9 or Trial 10, significant beta weights for gain are seen again.

The reasons for the lack of increased predictability in the intermediate phase are not entirely clear. It is possible that the spatial reversal task is characterized by two psychologically distinct phases which are masked in the intermediate trials. To determine the plausibility of this interpretation, the gain scores between each trial and each succeeding trial were correlated with the criterion. With one important series of exceptions, gain did not correlate with the criterion so that the increase of predictability could only be derived with gain as a suppressor variable. However, gain from Trial 4 to each succeeding trial did correlate significantly ( $r$ 's = .39, .39, .40, .49, .45, .45;  $p$ 's < .01), indicating that gain after Trial 4 was associated with less efficient performance on the criterion. Apparently, in the intermediate trials, some persons were still mastering the skills from the first stage of practice, while others were already headed toward mastery of the task. Gain in the intermediate stages did not reflect the differential aptitude of the learners in the spatial reversal task.

To explore the possibility that there may be population differences with respect to the improvement of prediction using gain scores,



the subjects were sorted into two groups. Subjects whose criterion performance could be predicted better when gains were added to initial status formed one group ( $N = 27$ ) and subjects for whom gain either made no difference or decreased predictability formed the other group ( $N = 18$ ). Trial means and variances were computed separately for the two groups. None of the comparisons between means and variances reached significance at the .05 level.

### *Discussion*

The most striking finding from the spatial reversal data is that the utility of gain in prediction is highly dependent on the stage of practice over which gain is computed. Gain was found to increase the predictability of the criterion task significantly when computed between early trials or between the late trials, but added nothing when computed over the intermediate trials. There was some indication that two distinct processes may be confounding the meaningfulness of gain during the intermediate trials because gain after the termination of the rate process had a direct rather than a suppressor relationship with the criterion. These results for the intermediate trials are consistent with previous research indicating that different processes are involved in different stages of practice (Dunham et al., 1968), and with the expectation that observed gain between two points does not necessarily parallel asymptotic level of aptitude. A simple rate measure does not differentiate an individual with extremely underdeveloped potential from an individual whose potential is only slightly underdeveloped because of the impossibility of determining if the individual is at an ascending or descending rate phase on his/her hypothetical ability curve.

The results also indicate the potentiality of molar correlation analysis to select the optimum measurement points for gain by internal criteria. The pattern of rate and terminal processes governing intertrial consistency led in this study to the designation of three distinct stages of practice. Theoretically, the maximum contribution of gain to predictability would be from a stage in which there are individual differences in rate of change which are not accounted for by their asymptotic level, to a stage which reflects the asymptotic level. Gain from trials with a strong rate process, Trial 2 and Trial 3, to the beginning of the terminal process, yielded large and significant increases in predictability. Less clearly anticipated, but mirrored by the patterns of intertrial consistency, was the predictability between late trials but not between intermediate trials.

Differences in trial means and variability were not found between subjects differing with respect to the improvement of predictability by the addition of gain scores. Very likely, the factors that will differentiate the group which increased in predictability by using gain scores will be extrinsic variables, such as age, race, and SES, rather than individual differences intrinsic to status measurement.

### *Conclusion*

The spatial reversal data indicate that the approach suggested here to distinguish aptitude from ability psychometrically has some feasibility. It was shown that predictability could be increased by adding an index of modifiability to initial status from a test-intervention-retest sequence. However, the data also indicate the crucial importance of the stage of practice or degree of intervention over which modifiability is computed. The degree of success with which aptitude can be distinguished from ability depends directly on the intervention or amount of practice which is selected. Equally important, but less obvious, is the probable influence of the degree of intervention in determining which population will show the most modifiability.

As shown in the spatial reversal data, there may be more than one stage of practice or degree of intervention which will provide increased predictability. Although not tested in this study, it is likely that different populations will be favored by modifiability measures taken from different stages of practice. If sub-cultures can be said to differ with respect to favorableness of the environment to the development of a given ability, then the average initial status points of these populations on their aptitude-ability curves will vary. One population may be at a very low point on this curve due to an extremely disadvantageous environment while another is at a mid-range point. The difficulty is, as previously discussed, that the average rate of change between two points does not reflect whether instantaneous rate is increasing or decreasing. Thus, if a low degree of intervention is chosen, the disadvantaged population may show a slow rate of change. There would be no way of knowing if this were due to being at the end of the curve (where rate decreases) or at the beginning (where rate increases).

In applying this aptitude-ability approach to complex tests, there are some other issues that must be resolved. These issues involve the reliability of a gain score. Many of the difficulties surrounding the use of change scores are avoided in the aptitude-ability model, since gain is a predictor rather than a dependent variable. As with any

other predictor variable, gain may have both true and error components. However, other paradoxes are unresolved. For instance, although classical test theory has assumed that errors of measurement are the same at each score level, it is probably true that scores at the low end of the distribution are more unreliable than those at the higher end. This means, then, that gain will be directly correlated with unreliability. A successful distinction of aptitude and ability may necessitate some reformulation of measurement theory or practice with respect to gain.

These considerations have important implications for research on the aptitude-ability model. The most basic implication is that research which has been traditionally conducted only during test validation must now be conducted during test development. Individual differences in change over practice must be studied by sensitive techniques, such as molar correlation analysis (Jones, 1970a), so that stages of practice can be determined. The influence of individual differences in extrinsic factors, such as population characteristics and other demographic factors, on change in different stages of practice must be studied simultaneously so that the intervention which provides the most useful modifiability index can be selected. Furthermore, it may be possible to deal with some of the paradoxes surrounding gain scores during test development, by designing tests specifically for measuring change.

Cronbach and Furby (1970) have considered the selection of individuals on the basis of residualized gain such as suggested here, to be unclear as to purpose. That is, it is difficult to determine if the unexpected gain was accidental, due to underestimation by the pretest or overestimation by the post-test. Thus, it is unclear as to how these individuals should be differentially treated. However, the problem noted by Cronbach and Furby is actually an empirical question: will the use of residualized gain scores lead to increments in predictability? If the answer to this question is affirmative, then it can be assumed that high unexpected gains are due to underestimation by the pretest. If some of the difficulties involved in measuring the modifiability of complex traits can be remedied, a successful psychometric distinction between aptitude and ability will have importance both theoretically and in application. Special educational resources and remedial training programs can be selectively applied to those who would profit the most.

#### REFERENCES

- Anatasi, A. *Differential psychology*. New York: Macmillan Company, 1958.

- Cronbach, L. J. and Furby, L. How we should measure change—or should we? *Psychological Bulletin*, 1970, 74, 68–80.
- Cronbach, L. J. and Snow, R. E. Individual differences in learning ability as a function of instructional variables: Final report to U. S. Office of Education, Contract No. OEC 4-6-0612 69-1217. Stanford: Stanford University, March, 1969.
- Dunham, J. L., Guilford, J. P., and Hoepfner, R. Multivariate approaches to discovering the intellectual components of concept learning. *Psychological Review*, 1968, 75, 206–221.
- Herrnstein, R. I.Q. *Atlantic Monthly*, 1971, 228, September, 43–64.
- Jensen, A. R. How much can we boost I.Q. and scholastic achievement? *Harvard Educational Review*, 1969, 39, 1–72.
- Jones, M. B. A two-process theory of individual differences in motor learning. *Psychological Review*, 1970, 77, 353–360(a).
- Jones, M. B. Differential processes in acquisition. In E. A. Bilodeau (Ed.), *Principles of skill acquisition*. New York: Academic Press, 1969.
- Jones, M. B. Practice as a process of simplification. *Psychological Review*, 1962, 69, 274–294.
- Jones, M. B. Rate and terminal processes in skill acquisition. *American Journal of Psychology*, 1970, 83, 222–236(b).
- Lee, E. S. Negro intelligence and selective migration: A Philadelphia test of the Klineberg hypothesis. *American Sociological Review*, 1951, 16, 227–233.
- Woodrow, H. Factors in improvement with practice. *The Journal of Psychology*, 1939, 7, 55–70.
- Wright, B. and Panchapakesan, N. A procedure for sample free item analysis. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1969, 29, 23–48.



## A MEASURE OF THE AVERAGE INTERCORRELATION

EDWARD P. MEYER<sup>1</sup>

University of Chicago

Bounds are obtained for a coefficient proposed by Kaiser as a measure of average correlation and the coefficient is given an interpretation in the context of reliability theory. It is suggested that the root-mean-square intercorrelation may be a more appropriate measure of degree of relationship among a set of variables.

GIVEN a set of statistical variables, the nature of the linear relationships among the variables can be summarized by the matrix of intercorrelations of the variables. A researcher may, on occasion, wish to obtain some measure of the "average" intercorrelation, either as an estimate of some common population value or as an indication of the degree of relationship among the variables as a group. Kaiser (1968) has presented rationale for using as a measure of average correlation a coefficient, gamma ( $\gamma$ ), which is a function of the largest eigenvalue of the correlation matrix and the number of variables. The purpose of this paper is twofold: first, to show that gamma is bounded numerically by two traditional measures of average correlation with equality obtaining under certain conditions and, secondly, to relate gamma to an estimate of average correlation obtained by applying the Spearman-Brown formula to a generalization of Cronbach's coefficient alpha. It is hoped that these developments will shed some light upon the psychometric properties of gamma and suggest possible cautions with regard to interpretation of the coefficient.

---

<sup>1</sup> I am most indebted to Professor Henry F. Kaiser for many helpful criticisms of an early draft of this paper.

Send reprint requests to The University of Chicago Drug Abuse Rehabilitation Program, 1440 South Indiana Avenue, Chicago, Illinois, 60605.

Copyright © 1975 by Frederic Kuder

*Kaiser's Coefficient Gamma*

Let  $P = [\rho_{ij}]$  denote a correlation matrix of order  $n$  and let  $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$  denote the ordered eigenvalues of  $P$ . Then the average correlation measure proposed by Kaiser (1968) is the quantity

$$\gamma = (\lambda_n - 1)/(n - 1). \quad (1)$$

Since  $P$  is a correlation matrix,  $\lambda_n$  must lie in the interval  $[1, n]$  and hence, by definition (1), gamma must lie in the interval  $[0, 1]$ . It will first be shown that it is possible to obtain bounds on gamma which are tighter than the bounds 0 and 1.

*Lower Bound for Gamma*

Notationally, given the matrix  $P$  of intercorrelations of a set of  $n$  variables  $z_i$  and the  $n$ -component constant weight vector  $x = [x_1, x_2, \dots, x_n]'$ , define the function

$$\lambda(x) = x'Px/x'x, \quad x \neq 0. \quad (2)$$

It is well-known (e.g., Bellman, 1960) that

$$\lambda_1 \leq \lambda(x) \leq \lambda_n, \quad x \neq 0, \quad (3)$$

with equality obtaining on the left or on the right in (3) if and only if  $x$  is a characteristic vector of  $P$  associated with  $\lambda_1$  or  $\lambda_n$ , respectively.

Setting  $x = 1 = [1, 1, \dots, 1]'$  in (2) yields

$$\lambda(1) = 1 + (n - 1)\bar{p}, \quad (4)$$

where

$$\bar{p} = 1/n(n - 1) \sum_{i \neq j} \sum \rho_{ij} \quad (5)$$

is the arithmetic mean of the correlations  $\rho_{ij}$ . It then follows from (3) and (4) that

$$\bar{p} \leq \gamma, \quad (6)$$

with equality if and only if  $1$  is a characteristic vector of  $P$  associated with  $\lambda_n$ .

More generally, let  $\mathcal{E}$  denote the set of  $2^n$  distinct  $n \times 1$  vectors which can be constructed when the range of the constants  $x_i$  is restricted to the two values  $+1$  and  $-1$  and define

$$\bar{p}_{\max} = [\text{Max } \lambda(x) - 1]/(n - 1), \quad x \text{ in } \mathcal{E}. \quad (7)$$

It then follows from (2) and (3) that

$$\bar{p}_{\max} \leq \gamma, \quad (8)$$

with equality if and only if the vector  $x$  in  $\mathcal{E}$  yielding  $\bar{p}_{\max}$  is also a characteristic vector of  $P$  associated with  $\lambda_n$ . Inequality (8) simply states that  $\gamma$  must be greater than or equal to the maximum value of  $\bar{p}$  for all possible reflections of the original variables.

### Upper Bound for Gamma

It will be shown that an upper bound for  $\gamma$  can be obtained by application of the following lemma to the eigenvalues of  $P$ .

#### Lemma

Let  $\mathcal{A} = \{a_1, a_2, \dots, a_n\}$  be a set of  $n$  real numbers and define

$$\mu_n = 1/n \sum_i a_i \quad (9)$$

$$\sigma_n^2 = 1/n \sum_i (a_i - \mu_n)^2.$$

Then  $a_{\max}$ , the largest element in  $\mathcal{A}$ , must satisfy

$$a_{\max} \leq \mu_n + \sqrt{n-1} \sigma_n, \quad (10)$$

with equality if and only if the set  $\mathcal{A}$  contains one element  $a_{\max}$  and  $n-1$  equal remaining elements  $a_i \leq a_{\max}$ .

*Proof.*<sup>2</sup> Without loss of generality we can assume that  $a_n = a_{\max}$ . Let  $b_i = a_i - \mu_n$ ,  $i = 1, 2, \dots, n$ . Then  $b_n = b_{\max}$ ,  $\sum_{i=1}^n b_i = 0$ ,  $\sigma_n^2 = \sigma_b^2$ . Thus,

$$(n-1) \sum_{i=1}^{n-1} b_i^2 \geq \left( \sum_{i=1}^{n-1} b_i \right)^2 = (-b_n)^2 = b_n^2 \quad (i)$$

with equality if and only if  $b_1 = b_2 = \dots = b_{n-1}$ . Adding  $(n-1)b_n^2$  to both sides of (i), we obtain

$$nb_n^2 \leq (n-1) \sum_{i=1}^n b_i^2 = n(n-1)\sigma_b^2 = n(n-1)\sigma_n^2. \quad (ii)$$

Since  $a_n \geq \mu_n$ , (ii) becomes

$$a_n - \mu_n = |b_n| = \sqrt{b_n^2} \leq \sqrt{n-1} \sigma_n. \quad (iii)$$

The remark following (i) shows that equality holds if and only if

<sup>2</sup> The author is indebted to an unknown reviewer for suggesting this proof of the lemma.

$b_1 = b_2 = \dots = b_{n-1}$ ; or, equivalently, if and only if  $a_1 = a_2 = \dots = a_{n-1}$  (since  $b_i = a_i - \mu_a$ ). Q.E.D.

In terms of the eigenvalues  $\lambda_i$  of  $P$ , (9) becomes

$$\mu_\lambda = 1$$

$$\sigma_\lambda^2 = 1/n \sum_i (\lambda_i - 1)^2 = (n - 1) \bar{p}_{rms}^2, \quad (11)$$

where

$$\bar{p}_{rms} = [1/n(n - 1) \sum_{i \neq j} \sum_j \rho_{ij}^2]^{1/2} \quad (12)$$

denotes the root mean square intercorrelation of the  $n$  variables.

Making the substitution (11) in (10) yields

$$\lambda_n \leq 1 + (n - 1) \bar{p}_{rms} \quad (13)$$

or, equivalently,

$$\gamma \leq \bar{p}_{rms}, \quad (14)$$

with equality if and only if  $\lambda_1 = \lambda_2 = \dots = \lambda_{n-1}$ .

Since  $\bar{p}_{max}$  in (7) must be greater than or equal to zero and  $\bar{p}_{rms}$  in (12) must be less than or equal to one, it follows from (8) and (14) that  $\bar{p}_{max}$  and  $\bar{p}_{rms}$  provide bounds for gamma which are tighter than the bounds 0 and 1, respectively.

#### *Generalization of Coefficient Alpha and Average Intercorrelation*

If, without loss of generality, the  $n$  variables are assumed to be items of a test of length  $n$ , then it is possible to give an interpretation to gamma within the context of reliability theory.

First, define the functions

$$\alpha(w) = \frac{n}{n - 1} \frac{w'(\Sigma - D^2)w}{w' \Sigma w}, \quad w \neq 0, \quad (15)$$

and

$$\begin{aligned} \bar{p}_{\alpha,1}(w) &= \frac{\alpha(w)}{n + (1 - n)\alpha(w)} \\ &= \frac{1}{n - 1} \frac{w'(\Sigma - D^2)w}{w' D^2 w}, \quad w \neq 0, \end{aligned} \quad (16)$$

where  $\Sigma$  denotes the item (variable) covariance matrix and  $D^2 = \text{Diag}(\Sigma)$  is a diagonal matrix such that

$$P = D^{-1} \Sigma D^{-1}. \quad (17)$$



Equation (15) is a generalization of the Kuder-Richardson reliability coefficient, Cronbach's coefficient alpha, to the case where items are weighted unequally with weights  $w = [w_1, w_2, \dots, w_n]'$  and equation (16) is the corresponding generalization of the estimate of average intercorrelation obtained by applying the Spearman-Brown formula to (15) with  $1/n$  as the multiple of test length. The generalized coefficient alpha (equation [15]) has been considered by Mosier (1943) and by Lord (1958); the rationale underlying use of the Spearman-Brown formula to obtain an estimate of average interitem correlation which is independent of test length (equation [16]) can be found in Cronbach (1951).

Note that if the items (variables) are given equal weight i.e.,  $w = 1 = [1, 1, \dots, 1]'$ , then  $\alpha(1)$  in (15) reduces to Cronbach's coefficient alpha (Cronbach, 1951, equation [24]) and  $\bar{p}_{\dots}(1)$  in (16) corresponds to the estimate of  $\bar{p}$  that one would obtain by applying the Spearman-Brown formula to coefficient alpha with  $1/n$  as the multiple of test length (Cronbach, 1951, Equations [44] and [45]). When items (variables) are not weighted equally, formulas (15) and (16) provide more general, but conceptually equivalent, measures of reliability and estimated average intercorrelation respectively.

Making the change of variable  $y = Dw$  in (16),

$$\bar{p}_{\dots}(y) = [\lambda(y) - 1]/(n - 1), \quad y \neq 0 \quad (18)$$

and it is evident that

$$\text{Max } \bar{p}_{\dots}(w) = \text{Max } \bar{p}_{\dots}(y) = \gamma, \quad \begin{array}{l} w \neq 0 \\ y \neq 0, \end{array} \quad (19)$$

i.e., the maximum possible value of  $\bar{p}_{\dots}(w)$  is Kaiser's coefficient gamma.

Since, by the reciprocal relationship (16), the weight vector  $w$  which maximizes  $\bar{p}_{\dots}(w)$  is also the vector which maximizes  $\alpha(w)$ , it follows that gamma can be interpreted as the estimate of  $\bar{p}$  obtained by applying the Spearman-Brown formula to the maximum value of  $\alpha(w)$ , the generalized coefficient alpha, with  $1/n$  as the multiple of test length. In other words, if the item (variable) weight vector  $w$  is chosen so as to maximize  $\alpha(w)$ , then the corresponding  $\bar{p}_{\dots}(w)$  is also maximized and, as indicated in (19), the value of this maximum is gamma.

### *Interpretation of Gamma*

It should be evident from the original derivation of gamma (Kaiser, 1968) and the subsequent first centroid approximation (Cureton,

1971) as well as the result in this paper relating gamma to coefficient alpha, that, if one interprets correlation as degree of relationship, then gamma is an appropriate measure of average correlation among a set of variables only to the extent that one has a homogeneous (single factor) set of variables. To the extent that more than one factor is necessary to account for the correlations among the variables, gamma will tend to underestimate the true degree of relationship among the variables. In such cases,  $\bar{p}_{rms}$  may provide a more appropriate measure of the true degree of relationship among the variables. Since this paper has established that  $\gamma \leq \bar{p}_{rms}$ , with equality obtaining only in the "single factor" case with  $\lambda_1 = \dots = \lambda_{n-1} \leq \lambda_n$ ,  $\bar{p}_{rms}$  would appear to be a more general measure of degree of relationship among a set of variables.

As an example, taking the classic correlation matrix from Hotelling (1933):

$$\begin{bmatrix} 1.000 & .698 & .264 & .081 \\ .698 & 1.000 & - .061 & .092 \\ .264 & - .061 & 1.000 & .594 \\ .081 & .092 & .594 & 1.000 \end{bmatrix},$$

one finds  $\bar{p} = .278$ ,  $\gamma = .282$ ,  $\bar{p}_{rms} = .393$ . Since, for the Hotelling matrix,  $\lambda_1 = 1.846$ ,  $\lambda_2 = 1.465$ ,  $\lambda_3 = .521$ , and  $\lambda_4 = .167$ , the reason for the discrepancy between  $\gamma$  and  $\bar{p}_{rms}$  should be obvious.

## REFERENCES

- Bellman, R. *Introduction to matrix analysis*. New York: McGraw-Hill, 1960, 110-111.
- Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, 16, 297-334.
- Cureton, E. E. A measure of the average intercorrelation. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1971, 31, 627-628.
- Kaiser, H. F. A measure of the average intercorrelation. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1968, 28, 245-247.
- Lord, F. M. Some relations between Guttman's principal components of scale analysis and other psychometric theory. *Psychometrika*, 1958, 23, 291-296.
- Mosier, C. I. On the reliability of a weighted composite. *Psychometrika*, 1943, 8, 161-168.

## EFFECTS OF A CONFIDENCE WEIGHTED SCORING SYSTEM ON MEASURES OF TEST RELIABILITY AND VALIDITY

RICHARD C. PUGH AND J. JAY BRUNZA  
Indiana University

The validity of a confidence scored vocabulary test was investigated by demonstrating an increase in its reliability without changing the relative difficulty of the test items and without detecting any personality bias in the confidence scoring system. The reliability estimate of the vocabulary test increased from .57 using a traditional scoring system to .85 using a confidence scoring system. No significant interaction was found between the difficulty of the test items and the type of scoring system. Three personality measures failed to correlate significantly with the confidence scores of the vocabulary test.

OBJECTIVE multiple-choice tests have been designed to allow respondents to indicate their degree of confidence in each option of a given test item. Ahlgren (1969) has shown that confidence scoring systems are effective in improving the reliability of scores on objective multiple-choice tests. As the reliability increases, a change in the relative difficulty of test items may occur. If a change in relative difficulty should occur, additional shifts in the characteristics of the test would be suggested. Few studies can be found that report a comparison of relative item difficulties.

The purposes of this study were (a) to demonstrate an increase in the reliability of a multiple-choice vocabulary test using a confidence scoring system, (b) to determine if a change in the relative item difficulties occurs under a confidence scoring system, and (c) to assess the relationship of traditional and confidence scored forms of a vocabulary test with selected personality measures. These three purposes were related to the validity of confidence scored tests.

Typically, research studies concerned with the effects of confidence scoring systems utilize multiple-choice tests that measure subject matter achievement. In this study, a vocabulary test was selected to see whether reliability could be increased on a test that measured something other than classroom achievement. The vocabulary test also provided enough items to permit typical estimates of reliability within a reasonable length of time.

Measures of external control, risk taking, and cautiousness were obtained to assess personality factors that might relate to the vocabulary scores under the testing condition of confidence scoring. The internal-external scale delineated individuals according to differences in a generalized expectancy or belief in external control (Rotter, 1966). The Kogan and Wallach (1964) questionnaire assessed an individual's propensity for risk-taking behavior and the Gordon (1956) Personal Inventory measured the general trait of cautiousness.

### *Method*

#### *Subjects*

The Ss used in this study were graduate students enrolled in educational measurement courses in the Department of Educational Psychology at Indiana University. These were required courses for several degree programs and had students enrolled from various subject-matter fields. Four sections of the measurement courses were chosen to generate a sample of 84 Ss. The sample consisted of 55 females and 29 males. The mean age of the sample was 25.5 years.

#### *Procedure*

The Ss were administered a 48-item multiple-choice vocabulary test consisting of items from the I.E.R. Intelligence Scale (1946). Items were selected from each of five levels of the intelligence scale and were randomly assigned to one of two sections of the test. The two sections of the test were considered to contain alternate sets of items.

Selection of a scoring scheme was based primarily on two reports. deFinetti (1965) investigated various answering techniques and scoring methods in order to make an adequate appraisal of subjective probabilities related to confidence testing. Rippey (1970) reported a comparative study of different scoring functions for con-



fidence tests. In both studies evidence was presented that simple scoring systems yielded relatively favorable characteristics when compared to complex scoring systems. The simple scoring systems had the overall advantage of facilitating the communication of the scoring process to examinees.

The *Ss* expressed their degree of confidence in each option by assigning a number from 0 through 5 to each of the five options on the vocabulary items. The score on the item was the number *S* chose for the correct answer.

To demonstrate comparative reliabilities of the vocabulary test for differing scoring systems, three forms of the vocabulary test were created. The three forms (A, B, and C) consisted of the same items divided into two sections of 24 items each. The three forms differed only in the directions given to the examinees. The directions for Form A followed the traditional right-wrong format for both sections. The examinees were told that their scores would be the number of items answered correctly. The correct answers were given a weight of five for convenience in the analysis. The directions for Form B followed the confidence system for Section 1 and the traditional system for Section 2. The directions for Form C followed the confidence system for both sections. The three different forms were randomly assigned to the 84 *Ss*.

Prior to taking the vocabulary test all *Ss* were given a brief training session consisting of two parts. The first part was a presentation by *Es* of the confidence scoring system. The second part allowed *Ss* to use the confidence scoring system on practice test items.

The effect of the different directions for the three test forms was studied by comparing the difficulty of the test forms, sections, and individual items under the three sets of conditions. A three-factor analysis of variance was computed. Differences among the levels of the forms, the sections, and the items nested in sections and their interactions were determined.

The reliability of the vocabulary test forms was estimated for the two sections and the total test using analysis of variance (Hoyt, 1941). Relationships of the personality measures with the vocabulary test were assessed using product-moment correlation coefficients.

### *Results*

The relative difficulty of the three forms, the two sections, and the forty-eight items were assessed using analysis of variance. All three factors were considered fixed. Since all *Ss* responded to all items in both sections, repeated measures were assumed across sections and

items nested in sections. An analysis of variance is presented in Table 1.

No significant difference was found among the means of the three forms. Although the two sections were intended to be alternative sets of the vocabulary items, a significant difference ( $p < .05$ ) between sections was found. The section means indicated that the second section was more difficult than the first. No significant interaction was found between sections and forms. A significant difference ( $p < .01$ ) among the difficulty of items was found. This was expected since the items were selected from five levels of the I.E.R. Intelligence Scale. No significant interaction was found between the items and forms, indicating that the relative difficulty of the items did not differ significantly among the three forms.

Table 2 consists of selected characteristics of the sets of vocabulary test items.

Estimates of the reliability for Sections 1 and 2 for Form A were .48 and .42, respectively. An overall reliability estimate of .57 was found for Form A. Reliability estimates for Form B were .62 for Section 1 and .48 for Section 2. For Form C the reliability estimates were .70 for Section 1 and .78 for Section 2. An overall reliability estimate for Form C was found to be .85.

Product-moment correlation coefficients were computed between the three personality measures and each of the two sections along with the total score of the vocabulary test forms. None of the coefficients were statistically significant at the .05 level. The coefficients ranged from .17 to  $-.31$  but a coefficient of .37 was needed for a relationship to be statistically significant ( $df = 26$ ).

TABLE 1

*Results of an Analysis of Variance Produced by the Vocabulary Test Items Using Three Test Forms*

Source	df	MS	F
Between Ss			
Forms (F)	2	33.55	2.63
Ss within forms R(F)	81	12.77	—
Within Ss			
Sections (S)	1	17.02	4.07*
F $\times$ S	2	2.98	<1
Error	81	4.18	—
Items within sections I(S)	46	84.81	23.62**
Forms $\times$ I(S)	92	4.31	1.20
Error	3,726	3.59	—

\*  $p < .05$ .

\*\*  $p < .01$ .

TABLE 2

*Means, Standard Deviations and Reliability Estimates for Two Sets of Vocabulary Items Using Three Test Forms, and Coefficients of Correlation between the Sets of Vocabulary Items and Selected Personality Measures*

Form and Section	$\bar{X}$	SD	$r_{xx}$	Ext. Control $r_{xy}$	Risk $r_{xy}$	Caut. $r_{xy}$
Form A						
1	2.45	.61	.48	-.15	-.09	.15
2	2.21	.60	.42	-.18	-.18	.17
Overall	2.33	.49	.57	-.20	.05	.15
Form B						
1	2.05	.49	.62	-.31	.09	.01
2	1.99	.65	.48	-.20	-.07	-.05
Form C						
1	2.16	.54	.70	.09	-.05	.02
2	2.08	.66	.78	-.03	-.09	.00
Overall	2.12	.56	.85	.02	-.08	.01

### Conclusions

By using a confidence scoring system, the reliability of the vocabulary test was increased without apparently altering the relative difficulty level of the items. The estimates of reliability for the sections of the vocabulary tests under the confidence testing system were substantially higher (.62-.78) than reliability estimates of the same sections using the traditional scoring system (.42-.48). Pooling together the two sections of the vocabulary test, the reliability estimate for Form C using the confidence testing system was .85. This was substantially higher than the .57 reliability estimate for Form A using the traditional scoring system. The increase in length of time to answer the 48 items was from an average of 14 minutes for Form A to an average of 19.5 minutes for Form C, a factor of only 1.4 times. The increase in reliability could not be accounted for solely by an increase in length of time since the effective test length was more than 3 times.

No personality measures correlated significantly with any form of the vocabulary test. The vocabulary test was not found to have a personality bias since the personality measures were reliable enough to allow significant relationships to be found. Reliability estimates for the external control measure ranged from .66-.85, for risk taking .67-.77, and for cautiousness .76-.88 using analysis of variance. The validity of the vocabulary test was considered to be

improved based on the increase in reliability using a confidence scoring system, since no significant change in the relative difficulty of test items was found and no significant personality bias was detected.

## REFERENCES

- Ahlgren, A. Reliability, predictive validity, and personality bias of confidence-weighted scores. Paper presented at the American Research Association convention, Los Angeles, February 1969.
- deFinetti, B. Methods for discriminating levels of partial knowledge concerning a test item. *The British Journal of Mathematical and Statistical Psychology*, 1965, 18, 87-123.
- Gordon, L. V. *Gordon Personal Inventory*. New York: World Book, 1956.
- Hoyt, C. Test reliability estimated by analysis of variance. *Psychometrika*, 1941, 6, 153-160.
- I. E. R. *Intelligence Scale CAVD*. New York: Columbia University, 1946.
- Kogan, N. and Wallach, M. *Risk taking, a study in cognition and personality*. New York: Holt, Rinehart and Winston, 1964.
- Rippey, R. M. A comparison of five different scoring functions for confidence tests. *Journal of Educational Measurement*, 1970, 7, 165-170.
- Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, 1966, 80, (1, Whole No. 609) 1-28.



## A TEST FOR HOMOGENEITY OF REGRESSION WITHOUT HOMOGENEITY OF VARIANCE

DONALD H. McLAUGHLIN<sup>1</sup>

University of California, Berkeley

A likelihood ratio test procedure is described for testing for homogeneity of regression coefficients and population means across treatment groups without assuming homogeneous error variances. Maximum likelihood estimation equations are obtained assuming the criterion scores to be normally distributed, and a method for their solution is described. Examples of situation for which the procedure is appropriate are given, and extensions of the procedure are discussed. The procedure has been used on hypothetical and real data.

TESTS for treatment effects on the relation between a set of predictor variables and a criterion variable, such as tests for aptitude-treatment interaction (ATI) in educational settings, as well as treatment effects on population means, may be analyzed statistically, even though the residual variances differ significantly across treatment groups. The assumption of homogeneity of variance is employed in current methods of analysis primarily to reduce the complexity of the computation. The purpose of this paper is to develop a testing procedure based on a model including treatment effects on residual variances. The main difficulty in the procedure is the solution of simultaneous equations for maximum likelihood estimates of parameters of the model; and an application of Newton's method is described which has been tested and found to be quite efficient.

There are two methods for testing for ATI which the proposed procedure would replace. The most common method, which is

---

<sup>1</sup> Now at the American Institute for Research, Palo Alto, California.

normally used with a single predictor variable, is to divide each treatment group into two subgroups, on the basis of whether their predictor scores are above or below the median. The predictor is then treated as merely another factor in an ANOVA model. There are at least two disadvantages of this procedure. First, by disregarding predictor variance within the subgroups created, a great deal of information in the data is neglected, reducing the power of the test. Second, and more important for the application of the procedure to the case of multiple predictors, the number of subgroups necessary is an exponential function of the number of predictor variables, and unless the predictors are mutually independent in each treatment, there may be a great deal of difficulty obtaining observations in each subgroup. Also, the use of the simple ANOVA model assumes homogeneity of variances across treatments.

A more reasonable procedure has been developed (Gulliksen and Wilks, 1950) which is based on the estimation of regressions of the criterion on the predictors for each treatment. The Gulliksen-Wilks procedure is based on the assumption that the criterion score,  $Y_{jk}$ , for the  $j$ th subject in treatment group  $k$ , is normally distributed with a mean which is the sum of the population mean for that treatment and a linear combination of the subject's predictor scores and with a variance,  $\sigma_k^2$ . The procedure consists of three successive tests, and the problem which the procedure described below solves is that the tests are ordered and if one is "failed," successive tests are invalid. The first test is whether the variances are homogeneous across treatments. If homogeneity of variance is rejected, then the other two tests are not valid. The second and third tests are for homogeneity of regressions and of population means across treatments, and it is this pair of tests that is likely to be important to the researcher, even though the variances are not homogeneous. Therefore, using the Gulliksen-Wilks procedure, one risks the possibility that the outcome may be that his data are inappropriate for his hypotheses.

The procedure described below is based on the same model of the data as the Gulliksen-Wilks procedure. However, the second and third of the three tests are constructed in such a way that they do not depend on the outcome of the test for homogeneity of variance. The tests consist of comparisons of the likelihood of the data under the assumption that all three factors, means, variances, and regressions, may vary across treatment groups ( $H_{VVV}$ ) with three alternative hypotheses: that the regressions are homogeneous across treatments although the means and variances may vary ( $H_{VVB}$ ), that

the variances are homogeneous although the other parameters may vary ( $H_{VEV}$ ), and that the population means are homogeneous although the other parameters may vary ( $H_{EVV}$ ).

Before considering the details of the procedure, you may wish to evaluate the usefulness of this extension of the Gulliksen-Wilks procedure. There are at least two types of situations in which the assumption of homogeneity of variance is a burden. One is an ATI study in which unmeasured aptitudes intrude on the interaction with treatments. For example, a researcher might be interested in the interaction of two cognitive ability scores, such as verbal reasoning and spatial imagery, with two instructional methods, in teaching, say, experimental design. However, although attempts are made to equate the instructional methods on requirements of aptitudes other than the two of interest, the researcher may not wish to guarantee that he has measured all sources of aptitude effects. For example, if one of the instructional methods induces a greater anxiety during the study, if that anxiety is correlated with criterion scores, and if subjects differ in sensitivity to the anxiety-producing situations, then, assuming that anxiety was not measured, the results would show an apparent heterogeneity of error variance. Those treatments which induced greater anxiety would have an extra source of error variance. Therefore, the Gulliksen-Wilks test for ATI would not be valid.

The second type of situation where the advantages of the procedural independence of the three tests are apparent is in the analysis of covariance in which the treatment groups are actually naturally occurring groups. One may be interested in testing for differences in population means without regard to differences in other population parameters. For example, one may be interested in sex differences in mathematical problem-solving, corrected for amount of mathematical experience, even though the contribution of that experience to the criterion differs between sexes. If, in fact, there were sex differences in the predictor scores, and if the actual relation between experience and the criterion were not linear, then differences in the regressions between sexes would be expected. The ordinary procedure for analysis of covariance would, therefore, not be valid.

We turn now to a description of the alternative procedure. The data to be analyzed are observations of a single criterion score,  $Y_{jk}$ , and a vector,  $X_{jk}$ , of predictor scores for each of  $n_k$  subjects in the  $k$ th of  $m$  treatment groups. Each of the criterion scores is assumed to be an independent normally distributed random variable whose

mean is the sum of the population mean for that treatment,  $\mu_k$ , and a linear combination,  $\beta_k'X_{jk}$ , of the predictor scores for that subject, and whose variance is  $\sigma_k^2$ . Although not necessary, the scores are usually assumed to be interval level measurements rather than ratio scales, so the grand mean is subtracted from each score. The likelihood of the data given this model, which we have denoted  $H_{VVV}$ , is:

$$L(D | H_{VVV}) = \prod_{k=1}^m \prod_{j=1}^{n_k} (2\pi\sigma_k^2)^{-1/2} \cdot \exp(-(Y_{jk} - \mu_k - \beta_k'(X_{jk}))^2/2\sigma_k^2). \quad (1)$$

The likelihood of the data under the three alternative hypotheses to be considered,  $H_{VVE}$ ,  $H_{VEV}$ ,  $H_{EEV}$ , are obtained by replacing  $\beta_k$  by  $\beta$ ,  $\sigma_k^2$  by  $\sigma^2$ , and  $\mu_k$  by  $\mu$ , respectively.

Once the likelihood of the data under each hypothesis is calculated, using maximum likelihood estimates of the free parameters, the ratio,  $\lambda$ , of the likelihood of the data given the more constrained of two hypotheses to the likelihood given the less constrained model is calculated. It is well-known (Wilks, 1962) that  $-2 \log_e(\lambda)$  is asymptotically distributed as a chi-square variable with the number of degrees of freedom equal to the difference in the number of parameters constrained by the two models, if the more restricted model is true. Larger values of the statistic are to be interpreted as rejections of the more constrained of the two models on the basis of the data.

The problems in applying the procedure lie in the solutions of the equations for the maximum likelihood estimates of the free parameters. Simultaneous equations for the estimates are given in Table 1, for each of the four hypotheses. They can be seen to depend only on the sample moments of the criterion and the predictors in the treatment groups. In the cases of  $H_{VVV}$  and  $H_{VEV}$ , the solutions of these equations are straightforward: the estimates of the regression coefficients are obtained first, and those estimates are used in the equations for the estimates of the population means and error variances. These are the only two sets of equations which must be solved to apply the Gulliksen-Wilks procedure. For  $H_{VVE}$  and  $H_{EEV}$ , the hypotheses of equal regressions and of equal means, without assuming equal variances, the equations cannot be solved directly. However, they can be solved iteratively, by Newton's method. The amount of computation in such a solution is significant, but given the availability of a computer and intelligent choices of variables for iteration and starting values, the procedure is quite efficient, taking about three iterations to reach negligible error levels.



TABLE I  
Equations for Maximum Likelihood Parameter Estimates

Hypothesis	Variance, $\sigma_k^2 = :$	Mean, $\mu_k = :$	Regression Coefficients, $\beta_k' = :$
$H_{VVV}$	$\text{var}_k(Y) - \hat{\beta}_k' \text{cov}_k(X, Y)$	$\bar{Y}_k - \hat{\beta}_k' \bar{X}_k$	$\text{cov}_k(X', Y) [\text{cov}_k(X, X')]^{-1}$
$H_{VVX}$	$\text{var}_k(Y) - 2\hat{\beta}' \text{cov}_k(X, Y) + \hat{\beta}' \text{cov}_k(X, X') \hat{\beta}$	$\bar{Y}_k - \hat{\beta}' \bar{X}_k$	$\sum_{k=1}^m \frac{n_k}{\sigma_k^2} \text{cov}_k(X', Y) \left[ \sum_{k=1}^m \frac{n_k}{\sigma_k^2} \text{cov}_k(X, X') \right]^{-1}$
$H_{VZV}$	$\frac{\sum_{k=1}^m n_k (\text{var}_k(Y) - \hat{\beta}_k' \text{cov}_k(X, Y))}{\sum_{k=1}^m n_k}$	$\bar{Y}_k - \hat{\beta}_k' \bar{X}_k$	$\text{cov}_k(X', Y) [\text{cov}_k(X, X')]^{-1}$
$H_{ZVV}$	$\text{var}_k(Y) - \hat{\beta}_k' (\text{cov}_k(X, Y) + \bar{X}_k (\bar{Y}_k - \bar{\mu})) + (\bar{Y}_k - \bar{\mu})^2$	$\frac{\sum_{k=1}^m \frac{n_k}{\sigma_k^2} (\bar{Y}_k - \hat{\beta}_k' \bar{X}_k)}{\sum_{k=1}^m \frac{n_k}{\sigma_k^2}}$	$(\text{cov}_k(X', Y) + \bar{X}_k' (\bar{Y}_k - \bar{\mu})) [\text{cov}_k(X, X') + \bar{X}_k \bar{X}_k']^{-1}$

Note.—The variables  $X$  and  $\beta$  are vectors, and expressions in brackets are matrices. The expressions  $\bar{X}_k$ ,  $\bar{Y}_k$ ,  $\text{var}_k(Y)$ ,  $\text{cov}_k(X, Y)$ , and  $\text{cov}_k(X, X')$  refer to the sample moments of the data in the  $k$ th treatment group.

For  $H_{VVE}$ , a function  $F$ , of the vector of variances, is defined by:

$$F_k(\sigma^2) = \sigma_k^2 - (\text{var}_k(Y) - 2\hat{\beta}'(\text{cov}_k(X, Y) + \hat{\beta}'(\text{cov}_k(X, X')\hat{\beta}), \quad k = 1, \dots, m, \quad (2)$$

where  $\hat{\beta}$  is the function of  $\sigma^2$  given in the  $H_{VVE}$  row of Table 1. The maximum likelihood estimates of the  $\sigma_k^2$ 's are the roots of  $F$ . Initial values for the variances,  $\sigma_{(0)}^2$ , are set equal to the values obtained for  $H_{VVV}$ . Successive values of the variances are then obtained by the equation:

$$\sigma_{(t+1)}^2 = \sigma_{(t)}^2 - [J]^{-1}F_{(t)}, \quad (3)$$

where the  $(k, l)$ th element of the matrix  $J$  is the derivative of  $F_k$  with respect to  $\sigma_l^2$ , and is given by:

$$J_{kl} = \delta_{kl} - \frac{2n_l}{\sigma_l^2} (\text{cov}_k(X, Y) - \hat{\beta}'(\text{cov}_k(X, X')) \\ \cdot \left[ \sum_{h=1}^m \frac{n_h \text{cov}_h(X, X')}{\sigma_h^2} \right]^{-1} (\text{cov}_l(X, Y) - \text{cov}_l(X, X')\hat{\beta}),$$

where  $\delta_{kl}$  is one if  $k = l$  and zero otherwise.

For  $H_{EVE}$ , a function  $G(\mu)$  is defined by:

$$G(\mu) = \sum_{k=1}^m \frac{n_k}{\sigma_k^2} (\bar{Y}_k - \hat{\beta}_k'(\bar{X}_k) - \mu), \quad (4)$$

where  $\sigma_k^2$  and  $\hat{\beta}_k$  are functions of  $\mu$  as given in the  $H_{EVE}$  row of Table 1. The maximum likelihood estimate of  $\mu$  is then the root of  $G$ . The initial value of  $\mu$  is obtained by solving  $G(\mu) = 0$  with values of the variances and regressions estimated for  $H_{VVV}$ . Successive values of  $\mu$  are obtained by an equation analogous to (3), where the derivative of  $G$  with respect to  $\mu$  is:

$$\sum_{k=1}^m \frac{n_k}{\sigma_k^2} (2(\bar{Y}_k - \hat{\beta}_k'(\bar{X}_k) - \mu)^2 / \sigma_k^2 \\ + \bar{X}_k'(\text{cov}_k(X, X') + \bar{X}_k\bar{X}_k')^{-1}\bar{X}_k - 1).$$

Given maximum likelihood estimates of the free parameters, the likelihood ratios can easily be shown to depend only on the estimates of the population variances,  $\sigma_k^2$ . The chi-square statistics and their respective degrees of freedom are given in Table 2. The three tests are, of course, not statistically independent, because they are all based partially on the same statistics of the data. However, they are procedurally independent in that the validity of

TABLE 2  
Test Statistics

Test	$-2 \cdot \log_e(\lambda)$	df
$H_{VVE}$ VS. $H_{VVV}$	$\sum_{k=1}^m n_k \log_e (\hat{\sigma}_{k,HVV}^2 / \hat{\sigma}_{k,HVVV}^2)$	$p(m-1)$
$H_{VEV}$ VS. $H_{VVV}$	$\sum_{k=1}^m n_k \log_e (\hat{\sigma}_{k,HVV}^2 / \hat{\sigma}_{k,HVVV}^2)$	$m-1$
$H_{EVV}$ VS. $H_{VVV}$	$\sum_{k=1}^m n_k \log_e (\hat{\sigma}_{k,HVV}^2 / \hat{\sigma}_{k,HVVV}^2)$	$m-1$

Note.— $p$  is the number of predictor variables,  $m$  is the number of treatment groups, and the estimates of variance are the solutions to the equations expressed in the corresponding rows of Table 1.

each is independent of the outcome of the other tests, and they are orthogonal in the sense that the existence of treatment effects on any of the model parameters does not alter the values of the statistics for the other tests. For example, if an extra source of error variance is introduced into one of the treatment groups, which is not correlated with the predictor scores, then the test for homogeneity of regression will not be altered. On the other hand, each of the tests is based on all of the sample moments, so alteration of the sample moments would effect the outcome of all three tests.

There is a point to be made concerning the interpretation of homogeneity of regression in the context of apparent heterogeneity of error variance. The test for homogeneity of regression is sensitive to variation across treatments of the relative contributions of the set of predictors, rather than to variation of the total amount of predictability. In fact, a situation in which the error variance is homogeneous and the regression coefficients are all multiplied by a constant which is characteristic of each treatment, is theoretically indistinguishable, in terms of the model used, from a situation in which the regression coefficients are homogeneous and the error variances differ. However, this ambiguity is not of great importance in ATI designs because the researcher is usually looking for just those regression differences for which the test is sensitive: for differences in the relative contributions of different aptitudes to performance in different treatment conditions.

There is one extension which is simple to implement, however. It is based on the fact that the estimation procedures for any set

of treatment groups are independent of the procedures for other treatments, assuming that there are no parametric constraints between different sets of treatments. For example, in a two factor experiment, the parameters can be estimated at each level of one factor independently of their values at other levels of that factor. The likelihood of the data of several sets of treatment groups is just the product of the likelihood in each set, so tests can be made between the alternative hypotheses that the parameters of the model are homogeneous within sets of treatments versus that they are not so homogeneous. In the two factor experiment, this means that the effects of each treatment factor can be examined without regard to the other factor.

This extension has been included in the FORTRAN IV computer program used to test the procedure. That program is available upon request, although it consists of little more than an expression of the equations in this paper.

#### REFERENCES

- Gulliksen, H. and Wilks, S. S. Regression tests for several samples. *Psychometrika*, 1950, 15, 91-114.  
Wilks, S. S. *Mathematical statistics*. New York: Wiley, 1962.



## CONVERGENT AND DIVERGENT MEASUREMENT OF CREATIVITY IN CHILDREN<sup>1</sup>

WILLIAM C. WARD

Educational Testing Service

Fourth through sixth grade children were given two types of creativity measures—divergent measures in which the child named all the ideas he could that met a simple requirement, and convergent measures, adaptations of Mednick's Remote Associates Test, in which he attempted to find one word that was associatively related to each of three others. Divergent and convergent measures shared little variance, and the latter were strongly correlated with IQ and achievement. Moreover, convergent items requiring production of the correct association were strongly related to items requiring only recognition. It was argued that in children Remote Associates performance depends on individual differences other than the size of the associative repertoire.

MEDNICK conceptualized the creative process in associative terms, seeing it as involving "the formation of associative elements into new combinations which either meet specified requirements or are in some way useful" (Mednick, 1962, p. 221). Individual differences in creativity were seen as depending on differences in the number and relative strength of associates the individual has available that are relevant to a problem. This formulation is schematic—what constitutes an element is not explicated, and several processes by which elements can come into association are mentioned but not

---

<sup>1</sup>This research was supported by research grant 1 P01 HD01762 by the National Institute of Child Health and Human Development to Educational Testing Service. Appreciation is due to Mrs. Sadie Mitchell, Principal, and to the teachers of the Paul L. Dunbar Elementary School for their cooperation; to Miss Patricia Warren, Miss Henrietta Gallagher, and Miss Suzanne Taweel for assistance in data collection and analysis; and to Drs. A. Harvey Baker and Nathan Kogan for critical reviews of the manuscript.

explained. It is also limited in scope—the discussion is focused on the associative substrate required for creativity, and does not include description of the control processes, e.g., personality and motivational variables, that must influence whether and how creativity is manifested in a problem situation. Nonetheless, it has been highly influential, since it provides a link between a highly complex phenomenon and simpler, ostensibly better understood, associative processes.

Two kinds of creativity measures have been rationalized in terms of this scheme. One of these, Mednick's Remote Associates Test (RAT), provides the subject with three words and requires that he find an additional word which is associatively related to all of those given (Mednick, 1962). For example, he is given *surprise*, *line*, and *birthday*; the solution word is *party*. Subjects presumably attempt to solve the problem by scanning their networks of associations to each of the problem elements and testing whether one of the resultant associations is common to all the networks. The creative subject has more associations available and therefore is more likely to find the one which satisfies the requirement.

Wallach and Kogan (1965) measured the extensiveness of the associative repertoire much more directly. Their subjects were asked for all the ideas they could give that met a simple problem requirement; for example, to name uses for an object, such as a shoe. Here the creativity measures were the number of relevant ideas, and the number of such ideas which were unique, given to each task. They noted that these two measures were likely to be related to one another: "... it is quite possible that more frequent associations will occur earlier and more unique associations later in a sequence, so that individuals who are able to produce a larger number of associations also should be able to produce a greater number of unique ones" (Wallach and Kogan, 1965, p. 14).

The two types of tests differ in that the RAT is convergent in form, requiring the production of a single predetermined solution to each problem, while the Wallach-Kogan measures are divergent, requiring many solutions. Nonetheless, both are rationalized as tests of the size and scope of the supply of associations the subject is able to generate given a simple problem; they differ in the directness of the test of this supply, not in the hypothesized continuum of individual differences that is under examination. It is worth examining, therefore, whether the two kinds of performance are related to one another. If they are substantially intercorrelated, it would provide evidence that the number of associations available

is indeed an important individual difference variable underlying performance on the Remote Associates Test. If not, Mednick's explanation of RAT performance—though not necessarily the usefulness of the test—can be called into question.

The present study tested the relationship of the two kinds of measures in fourth through sixth grade children. The Wallach-Kogan measures were designed for children of this age and required no important modification. The Remote Associates Test, however, was intended for adults. Two equivalent forms of the test were developed for use in this study, using some items taken from an unpublished children's version of the Remote Associates Test (Mednick and Mednick, 1962), plus a number of new items. Half the items in each form were presented as in the adult versions of the task, while half were given in a recognition format—each could be answered with one word from a list printed at the bottom of the test form. Use of this format served two purposes. First, it helped to assure that at least one part of the test would be of an appropriate difficulty level for children at each of the grade levels tested. Second, if both kinds of items should fall at a reasonable difficulty level for the children in this study, the interrelation of the two parts of the test would provide a further test of the degree to which the number of associates the subject has available is the crucial factor determining his level of performance. A recognition format should eliminate any differences dependent on the efficiency of memory search (McCormack, 1972), making the possession of the associative link and the ability to evaluate correctly its relevance the sole requirements for correct performance.

### *Method*

#### *Subjects*

Subjects were the 65 children, 26 males and 39 females, in one fourth, one fifth, and one sixth grade class of a predominantly black urban elementary school. Fourth grade Lorge-Thorndike IQ's, available on approximately half the sample, averaged 90.5 ( $SD = 13.2$ ).

#### *Measures*

Modified versions of two of the creativity measures developed by Wallach and Kogan (1965), the Uses and Pattern Meanings tests, were employed. In the first of these the child was asked to name uses for a common object; in the second, he gave possible interpretations of a simple abstract pattern. Each test consisted of

an example, followed by four test items; each test item was presented on a separate ruled page in a test booklet.

Two twenty-item forms of the Remote Associates Test were also employed. Each item consisted of three words, all associatively related to the same fourth word. Items were randomly assigned to forms and to order within forms. After instructions and four examples, the child was given one page containing 10 items in a recognition format. The 15 words listed at the bottom of the page included the answers to all 10 of these items. On a second page were presented 10 more items on which the child had to generate his own answers.

### *Procedure*

The tests were administered to intact classes in two sessions during the same week late in the school year. In Session 1, subjects were given the Pattern Meanings Test and then one form of the Remote Associates Test. In Session 2, they were given the Uses Test, followed by the second form of the Remote Associates Test. All testing was conducted by the same female research assistant; a male aide was present during the sessions, and the teacher sometimes remained in the room.

Administrative details were kept similar for all measures, so as to avoid, so far as possible, the introduction of method differences into the comparison of convergent and divergent measures. On the two divergent tests, labeled "What can you use it for?" and "What could it be?", the tester read through the instructions with the subjects, presenting an example item and eliciting responses from the class. The subjects then wrote down their ideas for each item. They were given five minutes per item, a time limit which was generous for most subjects. Children were told not to worry about spelling; the tester and the aide were available to help with wording if needed. The general testing atmosphere was businesslike—children were kept to the task, but with as little emphasis on time limits or on the evaluative aspects of the situation as was feasible.

Instructions and examples for the convergent measures, labeled "Related Words," were also read through by the tester and the subjects. Each item was then read aloud by the tester; the child had one minute to find or generate the answer and write it in a blank next to the three given words.

### *Scoring*

The Remote Associates Test items were initially scored according to a key containing the intended correct answers. Two judges



then examined those answers that had been scored wrong and, in a few cases, agreed that an answer not on the list was acceptable. The Uses and the Pattern Meanings tests were scored for number of ideas—the total number given, less only repetitions, incomprehensible responses, and those judged to be inappropriate. These tests are generally also scored for uniqueness, the number of acceptable ideas which are given by one child in the sample; but in previous studies uniqueness and number of ideas have been so highly correlated, frequently in the .80's for the two scores derived from the same test, that this score appears to provide little additional information (Ward, 1968, 1969).

### Results

Both the divergent and the convergent creativity measures showed substantial increases in mean level of performance from the fourth grade class to the two older ones, with little difference appearing between the latter two. Correlations among the measures were computed within each class and then averaged over classes, using Fisher's *r*-to-*z* technique. Correlations were also computed on scores standardized on class means and standard deviations; these coefficients did not differ systematically from those reported below, and are not presented.

In Table 1 are shown the intercorrelations of the two divergent creativity measures, the two forms of the convergent creativity measure, fourth-grade Lorge-Thorndike IQ, and the composite score

TABLE 1  
*Correlations among Creativity and Ability Measures*

	Pattern Meanings		Remote Assoc. A		Remote Assoc. B		IQ		Achievement	
	<i>r</i>	<i>N</i>	<i>r</i>	<i>N</i>	<i>r</i>	<i>N</i>	<i>r</i>	<i>N</i>	<i>r</i>	<i>N</i>
Uses—Number	.67***	65	.32*	65	.19	65	.15	32	.27	42
Pattern Meanings—Number			.33*	65	.34*	65	.39*	32	.41*	42
Remote Associates—Form A					.82***	65	.63***	32	.64***	42
Remote Associates—Form B							.50**	32	.62***	42
IQ									.75***	39

\*  $p < .05$ .

\*\*  $p < .01$ .

\*\*\*  $p < .001$ .

from the preceding spring's administration of the Iowa Tests of Basic Skills; IQ and achievement data were unavailable for some subjects. Similar matrices of correlations were also calculated separately for each sex; no systematic differences were found. Each type of creativity measure possessed a high degree of reliability across alternative tests. The two divergent measures, number of ideas on the Uses and on the Pattern Meanings tests, intercorrelated .67, while the two forms of the Remote Associates Test had an intercorrelation of .82 ( $p < .001$  in each case). However, the two types of test had only a minimal relation to one another; their intercorrelations ranged from .19 to .34, with three of the four coefficients significant at the .05 level.

In fact, the convergent and divergent measures shared little variance that was not also shared with IQ and achievement scores. Achievement scores may be the better indication of general ability level in this sample. They are more recent, the achievement tests having been given one year before the present testing, while IQ was tested while the child was in the fourth grade. Achievement had a moderate positive relation to divergent creativity measures ( $r$ 's of .25, n.s., and .41,  $p < .05$ ), but a strong correlation with convergent creativity ( $r$ 's of .64 and .62,  $p < .001$ ).

In Table 2 are shown the partial correlations among divergent and convergent creativity measures with achievement held constant. While all the correlations in the matrix were somewhat reduced by the removal of achievement variance, each type of creativity measure continued to show substantial internal consistency ( $p < .001$ ); and the correlations between divergent and convergent creativity were reduced to negligible magnitude. A similar analysis was done, partialling out IQ rather than achievement, for the 32 students having complete creativity and IQ data. As before, correlations within divergent and convergent creativity

TABLE 2  
*Correlations among Creativity Measures with Achievement Held Constant*

	Pattern Meanings	Remote Assoc. A	Remote Assoc. B
Uses—Number	.64*	.20	.03
Pattern Meanings—Number		.09	.12
Remote Associates—Form A			.70*

$N = 42$  subjects with complete data on all the above measures.

\*  $p < .001$ .

remained high (.75 and .73, respectively;  $p < .001$ ), while correlations between these two types of measures all failed to achieve statistical significance (average  $r = .17$ ; range from .03 to .31).

Product moment correlations among the recognition and production parts of the Remote Associates Test were also obtained. Within each form of the test, these two parts were highly correlated, with  $r$ 's of .63 for one form and .71 for the other. Across forms, the recognition scores correlated .65 and the production scores correlated .58 (all  $p$ 's  $< .001$ ). The two kinds of items, finally, showed equivalent relations to standardized achievement scores; for the sum of the scores on the recognition items over the two forms of the test, the correlation with achievement was .65; while for the sum over production items, it was .64 ( $p < .001$ ). Thus, there is no indication that the two kinds of items required different abilities from the subject.

### *Discussion*

It has been a common problem in creativity research that one investigator's measure of creativity turns out to be unrelated to another's. To some extent, this problem represents differences in the choice of the level at which creativity is operationalized (Taylor, 1959). The two types of measures studied here, however, have been presumed to be measures not only at the same level, but of the same process variable—the number of relevant associations the subject has available in simple problem situations. The Wallach-Kogan measures provide a direct assessment of this variable; and in this study, as in earlier work, these measures proved to possess both substantial reliability across alternative tests and discriminability from general intelligence and achievement measures.<sup>2</sup>

Remote Associates performance shared little variance with the divergent creativity measures, and therefore appears, contrary to Mednick's rationale for the test, to depend on variables other than the size of the associative repertoire. Moreover, the correlational similarity between recognition and production scores suggests that factors associated with the speed or efficiency of memory search for the relevant associate are not critical for performance. One

---

<sup>2</sup> Wallach and Kogan (1965) argued the importance of an evaluation-free testing context for creativity measurement. A definitive test of this proposition has not been made; however, the present results, along with data presented by Ward (1971), suggest that a group testing situation in which time limits are ample and evaluational cues are minimized is adequate for creativity assessment.

possible contributor to test performance might be individual differences in evaluative abilities (Frederiksen and Messick, 1959; Guilford, 1956)—as well as possessing the appropriate association, the subject must be able to decide that it is indeed appropriate.

In the present data, whatever abilities are responsible for Remote Associates performance appear also to contribute to IQ and achievement test scores. Similar findings have been presented in work with older children (Belcher and Davis, 1971; Warren, 1971) and with adults (Laughlin, 1967). A few studies with adults have found the test to measure something more than general intelligence; for example, showing a positive relation to incidental learning (Laughlin, 1967; Mendelsohn and Griswold, 1966). With children, however, it remains to be demonstrated that Remote Associates represents more than an unusual approach to the measurement of general intellectual ability.

## REFERENCES

- Belcher, T. L. and Davis, G. A. Interrelationships among three standardized creativity tests and IQ. Paper presented at the meetings of the American Educational Research Association, New York, February 1971.
- Frederiksen, N. and Messick, S. Response set as a measure of personality. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1959, 19, 137-157.
- Guilford, J. P. The structure of intellect. *Psychological Bulletin*, 1956, 53, 267-293.
- Laughlin, P. R. Incidental concept formation as a function of creativity and intelligence. *Journal of Personality and Social Psychology*, 1967, 5, 115-119.
- McCormack, P. D. Recognition memory: How complex a retrieval system? *Canadian Journal of Psychology*, 1972, 26, 19-41.
- Mednick, S. A. The associative basis of the creative process. *Psychological Review*, 1962, 69, 220-232.
- Mednick, S. A. and Mednick, M. T. Remote Associates Test; Elementary Grades Level. Copyrighted, 1962, S. A. Mednick.
- Mendelsohn, G. A. and Griswold, B. B. Assessed creative potential, vocabulary level, and sex as predictors of the use of incidental cues in verbal problem solving. *Journal of Personality and Social Psychology*, 1966, 4, 423-431.
- Taylor, I. A. The nature of the creative process. In P. Smith (Ed.), *Creativity*. New York: Hastings House, 1959. Pp. 51-82.
- Wallach, M. A. and Kogan, N. *Modes of thinking in young children: A study of the creativity-intelligence distinction*. New York: Holt, Rinehart and Winston, 1965.
- Ward, W. C. Creativity in young children. *Child Development*, 1968, 39, 737-754.



- Ward, W. C. Rate and uniqueness in children's creative responding. *Child Development*, 1969, 40, 869-878.
- Ward, W. C. Creativity test performance in young children. Paper presented at the meetings of the American Educational Research Association, New York, February 1971.
- Warren, T. F. Creative thinking techniques: Stimulating original ideas in sixth grade students. Paper presented at the meetings of the American Educational Research Association, New York, February 1971.



## LONGITUDINAL STUDIES OF RISK TAKING ON OBJECTIVE EXAMINATIONS

MALCOLM J. SLAKTER AND KEVIN D. CREHAN<sup>1</sup>

State University of New York at Buffalo

ROGER A. KOEHLER

University of Nebraska

Risk taking on objective examinations (rtooe) is defined as guessing when the examinee is aware of a penalty for incorrect responses (Slakter, 1967). Since rtooe measures can be obtained from Ss ostensibly taking aptitude or achievement tests, they provide psychologists with useful disguised measures of risk taking. Prior studies have indicated that rtooe is related to dominance-submission (Votaw, 1936), maladjustment (Sherriffs and Boomer, 1954), vocational choice (Ziller, 1957), curriculum choice (Slakter and Cramer, 1969), and perception of risk in military situations (Torrance and Ziller, 1957). Furthermore, it has been demonstrated that examinees low in rtooe tend to be penalized on test score (Hammerton, 1965; Sherriffs and Boomer, 1954; Slakter, 1968a; Slakter, 1968b; Votaw, 1936). These latter studies have shown that when Ss low in rtooe are forced to respond to all items, their average test score increases even though the usual penalty for incorrect responses is applied. Hence, we have evidence that rtooe confounds the aptitude or achievement being measured by the examination, and therefore rtooe concerns individuals involved with educational measurement.

PAST studies of the relationship of risk taking to age or sex have all been cross-sectional in nature. For example, Wallach and Kogan (1961) compared older subjects (mean age approximately 70) with college students on a hypothetical choice dilemmas instrument. They found that the older subjects were more conservative. In

<sup>1</sup> Now at West Virginia College of Graduate Studies.  
Copyright © 1975 by Frederic Kuder

studies of six to ten year old children, Kass (1964) found no age difference in gambling with pennies in a slot machine. However, Cohen (1960, chapter 5) examined nine-, twelve-, and fifteen-year old children in a risk taking situation with candy for prizes. He found that the nine year old children took greater risks than the 12 year olds, who in turn took greater risks than the 15 year olds. Finally, in a cross-sectional study of rtooe in grades five through eleven (Slakter, Koehler, Hampton, and Grennell, 1971), results demonstrated that students in grades five, six, and seven displayed higher risk taking than students in grades eight through eleven.

Cross-sectional studies examining the relationship of sex and risk taking have reported conflicting findings. For example, Wallach and Kogan (1959, 1961) have found no consistent sex differences in risk. Kass (1964), however, reported that boys selected greater risks than girls with the slot machines. On the other hand, using a decision-making task with candy as the prize, Slovic (1966) found a sex by age interaction. His results indicated no sex differences in younger children from ages six to ten, but with eleven- to sixteen-year olds greater risk was manifested by the boys. Specifically with rtooe, no sex differences were found in college students (Slakter, 1967; Slakter and Cramer, 1969), eighth grade females were found to be higher in rtooe than eighth grade males (Slakter, 1969), but the opposite was found for ninth grade students (Swineford, 1941). In a study of grades five through eleven, Slakter et al. (1971) reported a weak and inconsistent relationship between sex and rtooe.

All cross-sectional studies suffer from limitations (see Hilton and Patrick, 1970) due to possible cohort differences (i.e., different populations at the different grade levels) and cohort changes (e.g., dropouts). Therefore, it was decided to conduct longitudinal studies of the relationship between rtooe and age. The present study examines longitudinal data providing information on (a) the relationship between rtooe and age, (b) the age by sex interaction, and (c) the stability of rtooe.

### *Method*

The measure of rtooe was based upon the use of nonsense items, where a nonsense item is defined as one that has no correct (or best) answer, and no incorrect answer for the given population. Previous research (Slakter, 1967; Slakter, 1969; Slakter and Cramer, 1969; Slakter and Koehler, 1968) has established that five nonsense items embedded in five legitimate items yield Kuder-



Richardson formula 20 (KR-20) reliabilities in the vicinity of .80. In addition, since there is evidence that *rtooe* is a general trait across different types of examinations (Slakter, 1969), convenient synonym-antonym vocabulary items were employed in the measure. Ss were directed to indicate whether the words had the same or opposite meaning, and were informed of the penalty for incorrect responses. The following is an example of a nonsense item used in the measure:

7. *marnel*.....*mild*.

Since "marnel" is meaningless in the English language, the above item has no correct or best answer. Hence, any response (i.e., "same" or "opposite") is assumed to be an example of *rtooe* behavior; if the item is omitted, a lack of *rtooe* behavior is indicated. In order that age differences could be examined, the nonsense items which formed the basis of the *rtooe* measure were constructed so that they could be used at all grade levels; the legitimate items were selected to be appropriate for the particular grade level, and usually appeared at a single grade level. The *rtooe* score assigned to an S was the proportion of nonsense items attempted.

Ss for the study were all available public school students in grades five through eleven in a large village in western New York State. For the first testing in 1968 there were a total of 1,070 Ss, consisting of 522 males and 548 females. The number in each grade varied from 118 to 228. The tests were administered to the Ss in their own classrooms by their own teachers. The teachers were instructed as to standardized procedures of administration. The Ss were generally led to believe that they were taking another aptitude examination in their school's testing program. The tests were given as Part I of an "aptitude" examination on the same day to all classes in a given school, and within several days to the entire school system. The same procedure was repeated in 1970 to collect data after the passage of two years time. At this second testing, there were 1,049 Ss, with 536 males and 513 females. The number in each grade varied from 110 to 190.

We can classify the data into four sets, which are not necessarily mutually exclusive. We have (a) a set of cross-sectional data from 1968 (described previously in Slakter, et al. [1971]) designated as 68X, (b) a set of cross-sectional data from 1970 designated as 70X, (c) a set of unmatched longitudinal data which includes all the students tested each time and symbolized UL, and (d) a set of

matched longitudinal data which involves only those students tested in both 1968 and 1970, symbolized ML.

### *Results and Discussion*

Table 1 presents the KR-20 reliabilities for the rtooe measure. The values for the 68X data ranged from .68 to .86 with a median of .83; the values for the 70X data varied from .69 to .86 with a median of .82. Hence, the five-item rtooe measure had comparatively high internal consistency for grades five through eleven in both test administrations.

The mean rtooe scores for the cross-sectional and longitudinal data are presented by sex within grade in Table 2. The numbers in parentheses are the sample sizes. (The numbers of Ss in grades nine through eleven exceed the numbers in grades five through eight because of transfers into the system at the ninth grade. This cohort change represents more of a problem with interpretation of cross-sectional data than longitudinal data.) Note the close similarity between the cross-sectional and matched longitudinal means for both the 68 and 70 samples. This similarity indicates little cause for concern due to bias from selection or nonrandom loss of subjects.

Table 3 provides the mean rtooe differences over the two year time period by sex. The first column was calculated by subtracting 68X rtooe means (Table 2) separated by two year intervals; e.g., the difference for males from grades five to seven (.01) was found by subtracting the 68X mean for fifth grade males (.92) from the 68X mean for seventh grade males (.93). The second column was found in similar fashion from the 70X means. The values in the third column (ML) were found by subtracting the mean in the 68ML column in Table 2 from the appropriate mean in the 70ML column; e.g., the change for females eighth to tenth grade (-.13) was calculated by subtracting the 68ML mean for eighth grade females (.73) from the 70ML mean for tenth grade females (.60). Entries in the last column were calculated by subtracting the 68X mean from the appropriate mean in the 70X data; e.g., the mean difference for females in the ninth to eleventh category (-.06)

TABLE 1  
*KR-20 Reliabilities for Cross-Sectional Data*

Grade	5	6	7	8	9	10	11
68X	.78	.68	.85	.76	.86	.85	.83
70X	.82	.76	.82	.69	.86	.84	.84

TABLE 2

*Mean rtotoe by Sex within Grade (Sample Size in Parentheses)*

Grade	Sex	68X	70X	68ML	70ML
5	m	.92 (75)	.78 (63)	.93 (57)	
	f	.85 (60)	.67 (60)	.86 (42)	
6	m	.92 (63)	.92 (80)	.92 (50)	
	f	.87 (56)	.93 (64)	.89 (43)	
7	m	.93 (59)	.76 (72)	.93 (50)	.74 (57)
	f	.96 (71)	.85 (52)	.95 (53)	.87 (42)
8	m	.79 (48)	.66 (60)	.81 (37)	.66 (50)
	f	.73 (70)	.63 (50)	.73 (59)	.63 (43)
9	m	.72 (99)	.79 (94)	.72 (63)	.85 (50)
	f	.71 (129)	.82 (96)	.69 (85)	.86 (53)
10	m	.76 (85)	.69 (80)		.71 (37)
	f	.68 (80)	.60 (95)		.60 (59)
11	m	.61 (93)	.71 (87)		.73 (63)
	f	.66 (82)	.65 (96)		.66 (85)

resulted from the subtraction of the 68X mean for the ninth grade females (.71) from the 70X mean for the eleventh grade females (.65). Note that this last column provides the mean differences for the unmatched longitudinal data.

From an inspection of Table 3 we see that the mean rtotoe changes for the cross-sectional data differ from those of the longitudinal data, both matched and unmatched. For example, in the cross-sectional data we find essentially no age change in rtotoe over the eighth to tenth grade period, whereas the longitudinal data provide evidence of a decrease in rtotoe for this grade interval. These differences between the cross-sectional and longitudinal mean changes may be attributed to cohort differences or to cohort changes in the cross-sectional data.

TABLE 3

*Mean rtotoe Difference for Two Year Period*

Grades	Sex	68X	70X	ML	UL
5 to 7	m	.01	-.02	-.19	-.16
	f	.11	.18	.01	.00
6 to 8	m	-.13	-.26	-.26	-.24
	f	-.14	-.30	-.26	-.24
7 to 9	m	-.21	.03	-.08	-.14
	f	-.25	-.02	-.10	-.13
8 to 10	m	-.03	.03	-.10	-.10
	f	-.05	-.03	-.13	-.13
9 to 11	m	-.11	-.07	.01	.00
	f	-.05	-.18	-.03	-.06

Since the matched longitudinal data provides information on changes in the same people over a two year period, we assume that these data are the most informative. Repeated measures analyses at the .05 level on the ML data indicate: (a) significant decrease in mean *rtooe* for the periods sixth to eighth grade, seventh to ninth grade, and eighth to tenth grade, (b) a significant sex by age interaction for the fifth to seventh grade time period, with the males displaying a significant decrease in *rtooe*, but the females remaining constant and (c) no difference in *rtooe* for the ninth to eleventh grade period.

It was conjectured that perhaps the education process accounts for the decrease in mean *rtooe* by teaching the high risk takers to become more conservative. This conjecture was investigated by examining the bivariate scatterplots of the test-retest data for grades 6 to 8, 7 to 9, and 8 to 10. If the conjecture were true, we would expect to observe that many *Ss* high on the first administration would be low on the second. An examination of the scatterplots failed to uncover this relationship. Therefore, no evidence was found to indicate that the decrease in mean *rtooe* is caused by the education process. Perhaps, as Cohen (1960, p. 110) says: "... a relatively higher proportion of the younger ones may prefer to gamble because their hope (or psychological probability) of winning the prize is relatively greater than those of the older children." We will shortly consider the possibility that the decrease in mean *rtooe* may be attributed to developmental change.

To summarize the analyses of the ML data, there was a strong tendency for *Ss* to decrease in mean *rtooe* over grades 6 to 8, 7 to 9, and 8 to 10. There appeared to be little evidence for an age by sex interaction, except at the fifth to seventh grade period, and no evidence for sex differences in mean *rtooe*.

Test-retest reliabilities (stabilities) over the two year period for the *rtooe* measure were calculated from the ML data and are presented in Table 4. Note that the *rtooe* measure is extremely unstable for males until the ninth to eleventh grade period, while with the females the *rtooe* measure becomes somewhat stable at the sixth to eighth grade period and quite stable from the ninth to eleventh grade. The high KR-20's that we found with the cross-sectional data (Table 1) together with these low test-retest reliabilities indicate that *rtooe* tends to be a temporary characteristic of male students in grades five through nine (and perhaps longer), and in females from grades five through eight. However, *rtooe* appears to be a lasting characteristic for females from grade nine to eleven.



TABLE 4

*Test-retest Reliabilities (Stabilities) over Two Year Period (ML Data)*

Sex	Grade period				
	5 to 7	6 to 8	7 to 9	8 to 10	9 to 11
male	.08	.10	.20*	-.32*	.38*
female	.15	.34*	.34*	.37*	.72*

\* Significant at .05 level.

In observing that mean rtooe tends to stabilize (Table 3) at about the same time that rtooe appears to become a more lasting characteristic (Table 4), one might speculate that Ss do not have a developed concept of risk-taking on objective examinations until the ninth grade. Piaget and Inhelder (1951) have suggested that the probability concept is not acquired until after age eleven. While other researchers have provided evidence that the probability concept may be acquired before age eleven (e.g., Yost, Siegel, and Andrews, 1962), Hale, Miller, and Stevenson (1967) found that mastery of more sophisticated applications of probabilistic concepts may not appear until grade 9. Therefore, it seems reasonable to assume that any risk-taking characteristic might not become stable until the subjects had first mastered the concept of probability.

Psychologists interested in using rtooe as a disguised measure of risk-taking need to keep in mind that rtooe appeared to be a stable characteristic for females only for the ninth to eleventh grade period. For males, rtooe was unstable over all grade intervals studied. Individuals involved in educational measurement need to consider that: (a) whereas a particular student tends to be quite consistent in rtooe at a given point in time, rtooe tends to be unstable until at least grade 9—hence, the high rtooe student who is not penalized in an early grade (say sixth) may decrease in rtooe and be penalized in a later grade (say eighth) and vice versa, (b) test-retest stability for aptitude and achievement score will tend to be lowered by the lack of stability in rtooe, (c) the rtooe strategy in terms of maximizing average test score becomes poorer over grades five to nine, and (d) whatever rationale "do not guess" directions have, their use before grade nine may be difficult to defend since there is some doubt that students have mastered the concept of probability as it relates to rtooe.

In summary, mean rtooe tends to decrease over grades five to nine after which it appears to remain steady—at least until grade eleven. There was no evidence of a relationship between sex and rtooe, and evidence for an age by sex interaction only at the grade

five to seven period. Finally, there is evidence that rtooe is a temporary characteristic in the early grades, and does not become lasting for females until the nine to eleven grade period, and does not become lasting for males over the grade intervals studied.

## REFERENCES

- Cohen, J. *Chance, skill, and luck*. Baltimore: Penguin, 1960.
- Hale, G. A., Miller, L. K., and Stevenson, H. W. Developmental changes in children's concepts of probability. *Psychonomic Science*, 1967, 9, 229-230.
- Hammerton, M. The guessing correction in vocabulary tests. *British Journal of Educational Psychology*, 1965, 35, 249-251.
- Hilton, T. L. and Patrick, C. Cross-sectional versus longitudinal data: An empirical comparison of mean differences in academic growth. *Journal of Educational Measurement*, 1970, 7, 15-24.
- Kass, N. Risk in decision-making as a function of age, sex, and probability preference. *Child Development*, 1964, 35, 577-582.
- Piaget, J. and Inhelder, B. *La genese de l'idée de hasard chez l'enfant*. Paris: Presses Univer. de Frances, 1951.
- Sherriffs, A. C. and Boomer, D. S. Who is penalized by the penalty for guessing? *Journal of Educational Psychology*, 1954, 45, 81-90.
- Slakter, M. J. Risk taking on objective examinations. *American Educational Research Journal*, 1967, 4, 31-43.
- Slakter, M. J. The effect of guessing strategy on objective test scores. *Journal of Educational Measurement*, 1968, 5, 217-221. (a)
- Slakter, M. J. The penalty for not guessing. *Journal of Educational Measurement*, 1968, 5, 141-144. (b)
- Slakter, M. J. Generality of risk taking on objective examinations. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1969, 29, 115-128.
- Slakter, M. J. and Cramer, S. H. Risk taking and vocational or curriculum choice. *Vocational Guidance Quarterly*, 1969, 18, 127-132.
- Slakter, M. J. and Koehler, R. A. A new measure of risk taking on objective examinations. *California Journal of Educational Research*, 1968, 19, 132-137.
- Slakter, M. J., Koehler, R. A., Hampton, S. H., and Grennell, R. L. Sex, grade level, and risk taking on objective examinations. *Journal of Experimental Education*, 1971, 39, 65-68.
- Slovic, P. Risk-taking in children: Age and sex differences. *Child Development*, 1966, 37, 169-176.
- Swineford, F. Analysis of a personality trait. *Journal of Educational Psychology*, 1941, 32, 438-444.
- Torrance, E. P. and Ziller, R. C. Risk and life experience: Development of a scale for measuring risk taking tendencies. USAF PTRC Research Report, 1957, No. 57-23.

- Votaw, D. F. The effect of do-not-guess directions upon the validity of true-false or multiple-choice tests. *Journal of Educational Psychology*, 1936, 27, 698-703.
- Wallach, M. A. and Kogan, N. Sex differences and judgment processes. *Journal of Personality*, 1959, 27, 555-564.
- Wallach, M. A. and Kogan, N. Aspects of judgment and decision-making; interrelationships and changes with age. *Behavioral Science*, 1961, 6, 23-36.
- Yost, P. A., Siegel, A. E., and Andrews, J. M. Nonverbal probability judgments by young children. *Child Development*, 1962, 33, 769-780.
- Ziller, R. C. Vocational choice and utility for risk. *Journal of Counseling Psychology*, 1957, 4, 61-64.





## THE EFFECT OF DOUBLE STANDARDIZED SCORING ON THE SEMANTIC DIFFERENTIAL

JACK R. HAYNES  
North Texas State University

Two semantic differentials were administered to 200 college students and analyzed by two scoring methods, regular and double standardized. Factor analyses of these two scoring methods revealed different factor structures in both number of factors and patterns of loadings. Individual profiles were constructed on the 12 bipolar adjectives from both the regular and double standardized scoring and from factor scores based on these two scoring methods. Groups were formed from these profiles by Wards Hierarchical Grouping Technique. Comparisons of the group memberships also demonstrated that the different scoring methods reveal different patterns. The major difference between the two methods appeared to be due to interindividual variability being associated with regular scoring while double standardized scoring reveals intraindividual differences.

ALTHOUGH the semantic differential has been used in a great number of studies, most of these have utilized this technique to measure various theoretical concepts or to compare groups on response patterns. A few studies have concentrated on the nature and consistency of the factor structure (Darnell, 1966; Elliott and Tannenbaum, 1963; Howe, 1964; Levin, 1965) while others have examined the metric properties of the semantic differential (Heise, 1969; Green and Goldfried, 1965; Messick, 1957; Miller, 1956; Osgood, Suci, and Tannenbaum, 1957). Most references to scoring procedures have been summarized in *The Measurement of Meaning* (1957). In general raw scores on the scales have been used and standard *R* technique factor analysis has also been employed. Apparently very little has been done with the semantic differential

in the areas of grouping by profile similarity and the area of scoring effects on factor structure has also been limited.

Broverman (1961, 1962) has stated that the nature of the factor analytic approach and the scoring procedure should be influenced by the particular theoretical interest of the investigator. Broverman has pointed out the differences in the factor structure when the standard *R* technique procedures are used as contrasted with a double standardized data matrix. The effect of double standardization is to remove the between person variability from the correlations and thus the factor structure would reveal intraindividual or intrapsychic variations while the factor structure of the raw score data matrix would contain between person variability. If these effects are true, then grouping individuals on the basis of between person variability and on the basis of within person variability should produce different group memberships. The major purpose of this study was to investigate the effects of these two approaches on factor structure and on grouping by profile similarity in the semantic differential. A secondary purpose of this study was to compare the amount of error in grouping when raw scores versus factor scores are used. The following hypotheses were tested: (1) there will be a difference between the factor structure of a raw score data matrix and the factor structure of a double standardized data matrix, (2) group membership based on profile similarity of raw scores will be different from group membership based on profile similarity of double standardized scores, (3) profile grouping based on factor scores will produce less error than profile grouping on raw scores.

### *Method*

#### *Subjects*

Two hundred college students, 79 females and 121 males, from general psychology classes were administered the semantic differential on two concepts. The mean age was 20.14 years with a range from 17 years to 39 years. Thirty-one major fields were represented with 111 freshmen, 41 sophomores, 30 junior, 17 seniors, and 1 graduate student.

#### *Instruments*

Two concepts, myself and home, were scaled on 12 bipolar adjectives. The adjectives were selected from previous lists (Osgood,

Suci, and Tannenbaum, 1957) to provide representation of the three scales evaluation, activity, and potency and were scored on a seven point scale. The adjectives were: good-bad, beautiful-ugly, clean-dirty, large-small, heavy-light, strong-weak, active-passive, sharp-dull, fast-slow, kind-cruel, fair-unfair, and rugged-delicate.

### *Procedure*

The subjects were given the two concepts to be scaled with the order of presentation of concepts and the sequence of bipolar adjectives being randomized. This procedure was done to reduce any response set, and the reason for two concepts was to investigate the equivalence or stability of the results.

After the semantic differentials were completed, the following analyses were done on each concept separately: (1) a  $200 \times 12$  raw score data matrix was obtained yielding the raw scores for each individual on each of the 12 bipolar adjectives, (2) the data matrix was then converted to a double standardized data matrix by converting the scores in each column into  $z$  scores and then converting these standard scores into  $z$  scores in each row, (3) each of these matrices were factor analyzed by a principal axis method with unities in the diagonal. A Varimax rotation was applied to all factors with latent roots greater than one, (4) factor scores were computed for each subject from both the raw scores and double standardized scores, (5) Wards Hierarchical Grouping Technique (1963) was applied separately to each of the two sets of factor scores. A criterion of four groups, within each scoring technique, was arbitrarily chosen for further comparison, (6) a comparison of group membership was done by uncertainty analysis reduction to determine if the different methods yielded different groups based on profile similarity.

### *Results*

As shown in Tables 1-4, the rotated factor structures of the raw scores were different from those of the double standardized scores.

The results of the uncertainty reduction analysis in Table 5 reveals very low relationships between group memberships based on the different scoring methods.

As seen in Table 6, the least amount of group heterogeneity or error is found when factor scores are used with the greatest reduction occurring from factor scores based on the factor analysis of raw scores.

TABLE 1

*Rotated Factor Matrix for Raw Scores on the Concept Myself*

Variables	Factors		
	I	II	III
Good-Bad	.20	.02	.69
Beautiful-Ugly	.44	-.23	.41
Clean-Dirty	.50	-.02	.55
Large-Small	-.14	.78	.11
Heavy-Light	-.18	.81	.11
Strong-Weak	.27	.69	.02
Active-Passive	.73	.15	.21
Sharp-Dull	.73	.08	.21
Fast-Slow	.84	.01	-.03
Kind-Cruel	.03	.10	.86
Fair-Unfair	.07	.08	.80
Rugged-Delicate	.31	.65	-.12

*Discussion*

Inspection of Tables 1-4 reveal that the factor structures for the raw scores and the double standardized scores are different. They differ in both the number of factors and the pattern of loadings, which supports hypothesis number one. This finding would be in keeping with Broverman's contention that two different psychological pictures are obtained from these separate analyses. The analysis of the raw data yielded the usual factors of evaluation, potency, and activity, but when only intraindividual variability was analyzed, the results were different. MacAndrew and Forgý (1963) have criticized Broverman's findings on the basis that

TABLE 2

*Rotated Factor Matrix for Raw Scores on the Concept Home*

Variables	Factors		
	I	II	III
Good-Bad	.87	.04	.24
Beautiful-Ugly	.60	-.10	.40
Clean-Dirty	.64	-.02	.19
Large-Small	.07	.83	.02
Heavy-Light	-.11	.85	-.06
Strong-Weak	.34	.23	.68
Active-Passive	.24	.02	.81
Sharp-Dull	.26	.02	.79
Fast-Slow	.21	.08	.81
Kind-Cruel	.85	.03	.24
Fair-Unfair	.82	.01	.16
Rugged-Delicate	.02	.55	.25



TABLE 3

*Rotated Factor Matrix for Double Standardized Scores on the Concept Myself*

Variables	Factors				
	I	II	III	IV	V
Good-Bad	.01	-.06	-.75	.13	.17
Beautiful-Ugly	.17	.14	-.78	.04	-.14
Clean-Dirty	.23	-.10	.04	.75	.13
Large-Small	-.77	.19	.13	-.10	.04
Heavy-Light	-.81	.15	.16	-.12	.08
Strong-Weak	-.11	.42	.31	.01	.57
Active-Passive	.52	.17	.46	.09	-.04
Sharp-Dull	.15	.17	.19	.09	-.84
Fast-Slow	.60	.37	.08	-.14	-.26
Kind-Cruel	.02	-.78	-.02	.24	.08
Fair-Unfair	.07	-.87	.07	-.07	-.04
Rugged-Delicate	.07	.05	.20	-.82	.24

Broverman's method of factor extraction produced rotated factors, and when rotated factors extracted from the principle components *R* technique was used, the results were comparable. The present study, however, revealed distinct differences between the rotated factor structure of the raw scores and the double standardized scores even though a principal axis solution was used in both instances.

When the usual *R* technique was used on the semantic differential, the three scales of evaluation, potency, and activity were found. These factors would be descriptive of the meaning of the concepts for individuals on a normative basis while the factors extracted

TABLE 4

*Rotated Factor Matrix for Double Standardized Scores on the Concept Home*

Variables	Factors			
	I	II	III	IV
Good-Bad	-.61	.22	-.07	-.33
Beautiful-Ugly	.14	-.02	-.02	-.80
Clean-Dirty	-.21	.10	.13	-.60
Large-Small	.45	.50	.24	.33
Heavy-Light	.44	.39	.34	.43
Strong-Weak	.03	-.35	-.68	.24
Active-Passive	.26	.19	-.80	.00
Sharp-Dull	.11	-.75	-.13	-.07
Fast-Slow	.11	-.76	.08	.23
Kind-Cruel	-.77	.05	.11	.01
Fair-Unfair	-.77	.02	.12	.06
Rugged-Delicate	.31	.21	.41	.25

TABLE 5

*D Values from Uncertainty Reduction Analysis for Group Membership Similarity*

Score <sup>a</sup> Comparison	Concept	
	Myself	Home
1 and 2	.25	.21
3 and 4	.15	.17

<sup>a</sup> 1 = raw scores.

2 = double standardized scores.

3 = raw data factor scores.

4 = double standardized factor scores.

from the double standardized scores would be descriptive of the meaning of the concepts on an ipsative basis. Whether individuals score high or low on any of these latter factors would be determined by the relationships between their standings on the other ipsative factors.

The results in Table 5 also demonstrate a very low relationship between the profile similarities of the different scoring methods. This finding substantiates hypothesis number 2, for if there were no differences between the factor structures, then people should have similar profiles regardless of scoring method used, but this situation was not found in the present study.

The effect of score transformation can be demonstrated in an example with hypothetical subjects. Table 7 contains the raw scores and profile difference values for four hypothetical individuals. The *d* values are the summed squared differences between individual profiles. If the grouping is based on these values, Ss 1 and 2 would form one group while another group would be formed by Ss 3 and 4. Thus the grouping is greatly influenced by normative relationships or level of performance. This type of grouping would be appropriate when the investigator was interested in groups formed primarily from interindividual differences. When the scores in Table 7 were converted to double standardized scores, as shown in Table

TABLE 6

*Error Magnitude within the Hierarchically Formed Groups*

Type of Scores	Concept	
	Myself	Home
Raw Data	143.33	166.14
Double Standardized	81.31	80.28
R. D. Factor Scores	31.24	43.74
D. S. Factor Scores	77.41	75.51

TABLE 7

*Raw Scores and d Values for Four Hypothetical Subjects*

Subjects	Factors			d Values			
	I	II	III	1	2	3	4
1	7	5	6		14	48	62
2	4	7	5			46	34
3	3	1	2				14
4	1	4	1				

8, one group would be formed by Ss 1 and 3 while Ss 2 and 4 would form the other group. The absolute size of the *d* values would differ due to the magnitude of the measures in Table 8. The utilization of the double standardized scores would generate profile similarity based upon shape rather than level of performance. Thus within semantic differential data, double standardized scoring would yield different patterns of meaning for concepts than the raw score profiles.

Another approach would be to group individuals on the basis of both kinds of scores to achieve greater homogeneity with respect to level and shape. Grouping individuals on profile similarity and then looking for other common traits has certain advantages over the usual approach of comparing established groups on some instrument. This latter approach often leads to considerable variability within groups which may disguise the nature of important characteristics, for the basis for group membership is not made on psychometric similarity. Profile analysis, however, may reveal important sub-groupings within larger classifications.

The third hypothesis predicted less error for factor scores than for raw scores. The error values in Table 6 support this hypothesis. The reason for this error reduction would be due to proper weighting of the variables in factor scores, and since there are fewer factors than variables, less cumulative error when computing difference

TABLE 8

*Double Standardized Scores and d Values on the Scores in Table 7*

Subjects	Factors			d Values			
	I	II	III	1	2	3	4
1	1.0	-1.4	.4		11.76	.03	12.29
2	-1.2	1.2	0			11.63	.69
3	1.1	-1.3	.3				11.90
4	-.8	1.4	-.7				

values for the profiles. If the variables all had equivalent loadings on a factor, little difference would be found between factor scores and summing raw scores. If, however, there is considerable variability among the factor loadings, factor scores would be more accurate, for the appropriate contribution of each variable to the total would be achieved.

### Summary

When the factor structures and profile groups for raw scores and double standardized scoring were compared, differences were found which were directly related to the scoring procedure. Analyses based on raw scores tend to yield results related to interindividual variability while double standardized scoring yielded intraindividual or ipsative sources of variation.

### REFERENCES

- Broverman, D. M. Effects of score transformations in Q and R factor analysis techniques. *Psychological Review*, 1961, 68, 68-80.
- Broverman, D. M. Normative and ipsative measurement in psychology. *Psychological Review*, 1962, 69, 295-305.
- Darnell, D. K. Concept scale interaction in the semantic differential. *Journal of Communication*, 1966, 16, 104-115.
- Elliott, L. L. and Tannenbaum, P. H. Factor-structure of semantic differential responses to visual forms and prediction of factor scores from structural characteristics of the stimulus shapes. *American Journal of Psychology*, 1963, 76, 589-597.
- Green, R. F. and Goldfried, M. R. On the bipolarity of semantic space. *Psychological Monographs: General and Applied*, 1965, 79, 31.
- Heise, D. R. Some methodological issues in semantic differential research. *Psychological Bulletin*, 1969, 72, 406-423.
- Howe, E. S. Three-dimensional structure of ratings of exploratory responses shown by a semantic differential. *Psychological Reports*, 1964, 14, 187-196.
- Levin, J. Three mode factor analysis. *Psychological Bulletin*, 1965, 64, 442-452.
- MacAndrew, C. and Forgy, E. A note on the effects of score transformations in Q and R factor analysis techniques. *Psychological Review*, 1963, 70, 116-118.
- Messick, S. J. Metric properties of the semantic differential. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1957, 17, 200-206.
- Miller, G. A. The magical number seven, plus or minus two. *Psychological Review*, 1956, 63, 81-97.
- Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. *The Measurement of Meaning*. Urbana, Ill: University of Illinois Press. 1957.
- Ward, J. H. Hierarchical grouping to optimize an objective function. *American Statistical Association Journal*, 1963, 58, 236-249.



## ITEM-ANALYSIS OF JOURARD'S SELF-DISCLOSURE QUESTIONNAIRE-21

W. BARNETT PEARCE  
University of Kentucky

BERNIE WIEBE  
Freeman Junior College

Jourard's 21-item questionnaire was analyzed as an instrument for measuring the self-disclosure of 331 undergraduates at three universities. Alpha coefficients of internal consistency indicated that the SDQ-21 was acceptably reliable. Jourard's identification of the items as highly, lowly or moderately "intimate" was only partly supported by an analysis of items which discriminated between high and low disclosers. Although most disclosure was reported about lowly intimate and least about highly intimate items, three of the four best discriminators between high and low disclosers were judged by Jourard to be of moderate intimacy. Further, those items which discriminated best differed between male and female subjects.

Most questionnaires used in self-disclosure research closely resemble Jourard and Lasakow's (1958) instrument. Reliability and validity data summarized by Jourard (1971) support the use of questionnaires to measure self-disclosure.

One of the most convenient SDQ's is Jourard's (1971, p. 215) 21-item instrument. The purpose of this study was to determine (a) the reliability and discrimination values of each item; (b) whether items judged highly intimate by Jourard discriminate between high and low disclosers more than items judged lowly intimate; and (c) sex differences in disclosure.

### *Method*

Undergraduates (181 males and 150 females) at three Universities described their disclosure to their best friend, to a friend or to an acquaintance. Internal consistency coefficients were calculated for male and female Ss and total *N*. Correlation coefficients were computed between each item and the summated ratings for the total instrument. The discrimination value of each item was determined by observing the size of the *t*-value of the difference between the scores of the upper and lower thirds of the sample. Summed scores and discrimination values of each item were compared to identify sex differences. All computations were performed on an IBM 360/40 computer using a summated ratings scaling program SUMRAT9 (Brickner, Daucavage, and Kern, 1972).

### *Results*

Alpha coefficients of internal consistency (Cronbach, 1970) were .90 for the combined sample, .88 for male and .91 for female Ss. Pearson *r*'s between each item and the whole test were generally higher for items highest in discrimination power.

Table 1 reports the correlation coefficients between each item and the whole-test scores, and ranks by size of *t* the relative discrimination power of each item. Means and deviations for the total test and items at each intimacy grouping are given in Table 2.

Disclosure about items 17, 8, 5, and 3 discriminated most between high and low disclosers. Least discrimination power was found in items 7, 11, 13, and 19. Of the seven items in the high intimacy group, #3 and #6 were in the upper third in discrimination power for males, and #3 and #8 for females. Mean disclosure for high intimacy items was lower than that for low intimacy items. For males, 6 of the 7 low intimacy items (all but #1) were in the lowest third in discrimination power; items 7, 11, and 13 were in the lowest third for females. There was little difference in total disclosure for male and female Ss.

### *Summary*

Reliability scores, both for the total test and for each item, are acceptable, but Jourard's identification of intimacy levels is only partially supported by discrimination values. Different items discriminated high and low disclosers among male and female Ss, although total disclosure was comparable. These data suggest that this SDQ is an acceptable instrument but that intimacy designa-

tions do not necessarily identify items which distinguish high and low disclosers. Items in the upper third in discrimination value for both male and female Ss involved regretted acts, feelings of maladjustment and immaturity, bothersome habits and unhappiest moments.

TABLE 1  
*Item Rankings and Correlation Coefficients*

	Rankings by <i>t</i> -value*			Pearson <i>r</i>		
	Total <i>N</i>	<i>M</i>	<i>F</i>	Total <i>N</i>	<i>M</i>	<i>F</i>
1. Views on marriage roles ( <i>L</i> ) <sup>b</sup>	5	3	10	.70	.66	.77
2. Depression, anxiety and anger ( <i>M</i> )	17	11	17	.68	.65	.72
3. Most regretted acts and why ( <i>H</i> )	4	5	6	.71	.63	.78
4. Religious views and participation ( <i>L</i> )	11	19	9	.67	.56	.77
5. Feelings of maladjustment and immaturity ( <i>M</i> )	3	6	2	.71	.65	.76
6. Guiltiest secrets ( <i>H</i> )	13	4	20	.61	.62	.62
7. Views on politics ( <i>L</i> )	21	21	21	.56	.50	.61
8. Bothersome habits and reactions ( <i>M</i> )	2	2	3	.77	.73	.80
9. Dissatisfaction with opposite sex ( <i>H</i> )	12	18	7	.69	.59	.79
10. Erotic play and sexual lovmaking ( <i>H</i> )	14	8	14	.62	.62	.63
11. Hobbies and leisure time ( <i>L</i> )	20	20	18	.68	.55	.77
12. Happiest occasions in life ( <i>M</i> )	7	14	4	.76	.65	.86
13. Aspects of daily work ( <i>L</i> )	19	16	16	.69	.63	.75
14. Positive personal charac- teristics ( <i>M</i> )	16	7	19	.69	.66	.71
15. Persons most resented ( <i>H</i> )	8	13	11	.69	.59	.77
16. Sexual intimacies ( <i>H</i> )	10	12	12	.63	.56	.69
17. Unhappiest moments ( <i>M</i> )	1	1	1	.75	.69	.81
18. Music preferences and dislikes ( <i>L</i> )	9	15	8	.74	.61	.84
19. Personal goals ( <i>L</i> )	18	17	13	.65	.59	.71
20. Personal depression and hurt feelings ( <i>M</i> )	6	9	5	.75	.68	.85
21. Sexual fantasies and reveries ( <i>H</i> )	15	10	15	.61	.62	.62

Note.—The 21 items are abbreviated forms of those used in the SDQ published in Jourard's *Self-Disclosure*, p. 215.

\* For all *t* values,  $p < .05$ .

<sup>b</sup> Parenthetical labels refer to intimacy ratings. *H* denotes high intimacy; *M*, moderate intimacy; and *L*, low intimacy. (cf: Note above).

TABLE 2  
*Means and Standard Deviations*

	Total <i>N</i>	21-items <i>M</i>	<i>F</i>	7-item intimacy groups <i>H</i>	<i>M</i>	<i>L</i>
Mean total scores	45.97	46.43	45.42	14.02	15.27	16.68
Standard deviation	10.87	9.56	12.37	4.10	4.07	3.80

## REFERENCES

- Brickner, C., Daucusavage, J., and Kern, G. *SUMRAT9*, Computer Center: University of North Dakota, 1972.
- Cronbach, L. J. *Essentials of psychological testing*. New York: Harper and Row, 1970.
- Jourard, S. M. *Self-disclosure: An experimental analysis of the transparent self*. New York: Wiley-Interscience, 1971.
- Jourard, S. M. and Lasakow, P. Some factors in self-disclosure. *Journal of Abnormal and Social Psychology*, 1958, 56, 91-98.



## THE CLASSROOM BOUNDARY QUESTIONNAIRE: AN INSTRUMENT TO MEASURE ONE ASPECT OF TEACHER LEADERSHIP IN THE CLASSROOM

THOMAS L. MORRISON<sup>1</sup>

Department of Psychiatry  
University of California, Davis

Two studies were done to operationalize the concept of social system boundaries as applied to teacher control in classrooms. In Study 1 a multiple choice Classroom Boundary Questionnaire (CBQ) was developed to measure teacher preference for boundary control. The 25 items were found to load primarily on one factor and to have adequate split-half reliability (corrected  $r = .85$ ). In Study 2, observations in 32 4th through 6th grade classrooms found that observational measures of teacher boundary control behavior could be reliably recorded and were correlated with teachers' CBQ scores. Degree of boundary control preferred by the teacher on CBQ and the frequency of child-initiated boundary crossing events allowed in the classroom were negatively correlated ( $r = -.48, p < .01$ ).

MAINTAINING control over the behavior of children in their classrooms has long been an important practical problem for teachers. A major complaint of teachers about their preparation for teaching is that the issues of control and motivation are not sufficiently emphasized (Wright and Tuska, 1968). For many years, there was

<sup>1</sup> This paper is based on sections of a dissertation submitted to the Psychology Department of Yale University in partial fulfillment of the requirements for the PhD Degree. The author wishes to express appreciation to Dr. James C. Miller, chairman of the dissertation committee, and to Drs. Donald Quinlan and Claude Buxton, members of the committee. The research was supported by a Predoctoral Research Fellowship from the National Institute of Mental Health (5 F01 MH49563).

The author's address is Department of Psychiatry, University of California, 2315 Stockton Blvd., Sacramento, California 95817.

Copyright © 1975 by Frederic Kuder

a lack of research evidence on which to base statements about the effects of teacher control (Ladd, 1958a, 1958b; Sheviakov and Redl, 1956). Though the amount of research on classroom interaction has increased in recent years, much of the social psychological and educational research on teacher control in classrooms has remained disjointed and without firm grounding in psychological theory (Morrison, 1972).

Recently, writers have emphasized the possibilities inherent in viewing the classroom as a group or social system, and the teacher as the group leader or manager (Getzels and Thelen, 1960; Jensen, 1960; Jenkins, 1960; Roberts, 1971; Schmuck and Schmuck, 1971). In this context, the question of teacher control becomes one aspect of teacher leadership in the classroom group. This paper uses the concept of boundary, developed in Tavistock social systems theory (E. J. Miller, 1959; E. J. Miller and Rice, 1967), to conceptualize questions relating to teacher control of child behavior in classrooms.

### *Study 1: Questionnaire Development*

#### *The Concept of Classroom Boundaries*

In social systems theory, the concept of boundary is used to analyze interactions between and within groups. Boundaries, which can be physical but need not be, occur at points of discontinuity in space, time, or behavior. A discontinuity is a boundary if there is control or regulation of transactions across it (J. C. Miller, 1971). It is an important function of the management of an organization to regulate transactions across the boundaries between the organization and other social systems in its environment. The management must also regulate transactions across the boundaries that separate subsystems within the organization itself. All systems can be thought of as having a task that requires taking in materials from the environment, processing them, and distributing a product. In order for the task to be accomplished effectively, the management of the system must regulate the flow of the material as it passes across the boundaries from one processing system to the next. That is, the operations required for doing the task must be coordinated.

The classroom can be thought of as a complex social system in which each child is a subsystem and the teacher is the manager. The task of the classroom group is the production of learning among its members. As the manager of the classroom group, the teacher has to decide how to control the transactions among group members so that the task of learning is accomplished effectively and efficiently.

A classroom with high boundary control would be one in which the transactions among students and between a student and the rest of the class were carefully regulated by the teacher. In such a classroom, for example, children would not talk to other children, would not leave their seats to go to another part of the classroom, and would not make a statement to the entire class unless the teacher specifically initiated or approved such actions. In a classroom with low boundary control, on the other hand, children would be free to initiate conversations with other children and would have free access to facilities in the room. With regard to the sequence of work tasks, the teacher in a classroom with high boundary control would specify in detail the task to be done and the time when it was to be done. In a classroom where the time boundary was less controlled, children would have more choice when to do their work.

Most elementary school classrooms have little direct interaction with the surrounding environment, so that most issues of boundary control relate to internal boundaries. However, there are some decisions that the teacher must make about the relationship between the classroom and the larger system of the school. In particular, the teacher must decide about the conditions under which children can leave the classroom: whether children can decide to leave the room themselves (low boundary control), whether they must first seek permission, or whether they may leave only at times indicated by the teacher (high boundary control). In general, then, the question of boundary control in classrooms relates to the degree of constraint on children with regard to their use of time and space in the classroom and with regard to the kinds of interactions they can have with other members of the class. This paper reports on the development of a questionnaire designed to operationalize the concept of boundary control in classrooms.

### *Method*

A multiple choice questionnaire, the Classroom Boundary Questionnaire (CBQ), was constructed to measure the boundary conditions that teachers say should prevail in their classrooms. Thirty questions were written to present a brief boundary-related situation. Following each situation was a list of three or four possible subsequent courses of action. Respondents were asked to choose the alternative that most nearly represented the behavior preferred in their classrooms. The alternatives within each question varied along a continuum that reflected the extent of boundary control exercised by the teacher in the classroom. Items covered teacher preferences

for control over three kinds of boundaries: space boundaries, time boundaries, and behavior boundaries. Brief descriptions of these boundaries (and examples of questionnaire items) follow.<sup>2</sup>

### *Control of Space Boundaries*

A child's assigned seat can be thought of as a bounded space. Questions relating to space boundaries asked teachers for their opinions about conditions under which children ought to be allowed to cross the boundary around their seats. A child can cross this boundary by talking to another child or by leaving his seat. In some classrooms, children can cross this boundary whenever they wish; in others, they may do so only with the permission or at the instruction of the teacher. The following item is about the boundary around the child's seat:

After finishing his assigned seatwork,

- (a) a child should feel free to leave his seat in order to get materials from the classroom library or from another part of the room.
- (b) a child should be permitted to leave his seat only after requesting the teacher for permission to do so.
- (c) a child should stay in his seat until the teacher directs him to some other activity.

Another space boundary is the boundary between the classroom and the rest of the school. Some questions asked about conditions under which children were allowed to cross this boundary. The following is an example:

A child who wishes to leave the room during a class period (e.g., to go to the bathroom or get a drink of water)

- (a) should feel free to get up and leave quietly.
- (b) should be allowed to do so, but only after asking permission from the teacher.
- (c) should be told to wait until recess or another scheduled break period.

### *Control of Time Boundaries*

The classroom day is bounded at the beginning and the end by clear times for the start of school and for dismissal. Within those limits teachers can vary in how much freedom of choice they allow children in their use of class time. Questions about time boundaries

<sup>2</sup> Copies of the Classroom Boundary Questionnaire and information about scoring are available on request from the author.



attempted to ascertain whether teachers allowed children some latitude in choosing when to do work, or whether they structured the children's use of time more closely. The following question is an example:

With regard to seatwork assignments

- (a) a child should feel free to work on his assignments in whatever order he wishes, provided he gets all his work done by the end of the day.
- (b) when a particular seatwork lesson (e.g., Arithmetic or Spelling) is assigned, the child should work only on that lesson until the teacher directs the class to a new lesson.
- (c) a child should not have to turn in routine assigned work on time if he is working on something else that is interesting to him.

### *Control of Behavior Boundaries*

Some items were written to determine how broad a range of behavior was acceptable in the classroom, or how clear the boundary was between acceptable and unacceptable behavior. Teachers were asked what kinds of physical and verbal behaviors were "within bounds." The following question is an example:

With regard to language in class,

- (a) children should feel free to use the language that comes naturally to them, even if it is ungrammatical (i.e., colloquialisms, slang, "ain't," etc.). Swearing, however, should not be allowed.
- (b) children should feel free even to swear in class if this expresses how they feel.
- (c) children should watch their language in the classroom: they should be polite and try to use correct English.

### *Judges' Ratings*

Establishing a scoring system for the questionnaire involved verifying that the alternative choices within each question did vary along a continuum reflecting teacher control of the boundaries. Each alternative within each of the 30 questions was rated on a 9-point scale by 26 psychologists and social workers, producing 96 ratings for each rater. The end points of the scale used by the raters were labelled (a) Not permissive vs. Very permissive; (b) Not much freedom allowed vs. Very much freedom allowed; and (c) Very much structure imposed vs. Not much structure imposed. The raters

were instructed to assign a different value to each alternative within each question.

The raters showed high agreement. For each of the 96 alternatives, a mean rating was determined, and the analysis of variance procedure suggested by Winer (1962, pp. 124-132) was used to estimate the reliability of each mean rating. The resulting reliability figure was  $r = .99$ . One question that showed particular disagreement among raters was dropped at this time.

### *Subjects*

Sixty-two graduate students of education served as Ss. Of these, 17 were in three sections of a practicum course for teachers studying to be guidance counsellors and 45 were in two sections of a graduate course in child development. Eighteen were men, the mean age was 27.4 years ( $SD = 5.6$ ), and the mean number of years teaching experience was 2.9 ( $SD = 2.3$ ). All but six of the Ss had had actual teaching experience beyond student teaching, and all but three of the rest were employed as teachers at the time they answered the questionnaires.

### *Questionnaire Administration*

Class time was made available for *E* to describe his research and for class members to complete several questionnaires. *E* briefly discussed the lack of research on classroom discipline, and described his project as a study of "how teachers'" attitudes and opinions relate to their decisions about what kinds of things should be allowed to happen in classrooms." The teachers were then asked to complete four questionnaires, three of which<sup>3</sup> were characterized as "ways of measuring attitudes and opinions." They were told that the fourth questionnaire (the CBQ) was developed by *E* "to find out what teachers think should happen in classrooms." CBQ items were designed to refer to 4th, 5th, and 6th grade classrooms. Teachers who did not normally teach those grades were asked to respond in terms of what they would expect in their classrooms if they did teach at that level. Four persons in the counselling practica declined to participate.

<sup>3</sup> The personality questionnaires were the Marlowe-Crowne Social Desirability Scale (Crowne and Marlowe, 1964), the Tomkins Left-Right Scale (Tomkins, 1963), and the Miller Boundary Questionnaire (Miller, 1968).

### *Scoring System*

In scoring the questionnaire the mean ratings that had been derived for each alternative within each question were transformed into rank orderings. Thus, each alternative received a value ranging from 1 to 3 or 4 (depending on the number of alternatives within the question). The higher the score, the more control was represented by the item. A total boundary control score was assigned to each *S* by summing the value for each alternative chosen.

There were four questions on which 75% or more of the teachers chose the same alternative. These questions were dropped, leaving 25 questions contributing to the score. The possible range of CBQ scores extended from 25 to 81.

### *Results*

The CBQ produced a range of scores from 29 to 62, with a mean of 45.9 ( $SD = 8.8$ ). With items assigned randomly to halves, the CBQ had a split-half reliability, corrected by the Spearman-Brown formula, of  $r = .85$ . No test-retest reliability figure is as yet available.

CBQ score did not correlate with the teacher's age or amount of experience as a teacher. The mean CBQ scores of male and female teachers were not significantly different. The teachers in the counselling practica had a lower mean CBQ score than the rest of the teachers ( $t = 3.67, p < .001$ ).

Though no formal assessment of *S*'s reactions to the questionnaire was attempted, it appeared that most teachers enjoyed the CBQ and found it relevant and realistic. Some complained that the multiple choice format kept them from saying how they really acted in the situation, or that it did not allow for the richness, complexity, and multiple contingencies of real classroom interaction.

### *Factor Analysis*

The items of the CBQ were subjected to a factor analysis. In the unrotated factor matrix, all but four items loaded on the first factor with  $r > .35$ , and this factor accounted for 26.2% of the variance. The next two factors accounted for 7.9% and 7.5% respectively, and the percentage dropped below 7% after that. A varimax rotation with two and three factors retained did not yield factors that were distinguishable. These results were interpreted as supporting the expectation that the CBQ would measure a single factor, namely the teacher's preference for control of the boundaries in the classroom.

*Study 2: Questionnaire Validation*

Validation of the Classroom Boundary Questionnaire involved the development of observational procedures to measure actual teacher boundary control behavior in classrooms.

With regard to space boundaries, at least four situations relating to the boundary between a child and the surrounding social system can be identified: a child leaving his seat, a child having a conversation with another child or several children, a child talking aloud to the entire class, and child leaving the classroom. There are three possible ways for such boundary-crossing events to be initiated. The teacher can instruct the child to cross the relevant boundary (e.g., asking the child to get up and collect papers), the child can seek permission from the teacher before crossing the boundary (this is usually done by the child raising his hand), or the child can take the initiative and cross the boundary without referring first to the teacher. In the latter case, the teacher can either reassert control of the boundary by correcting the child's behavior (e.g., a simple reminder not to talk, a reprimand, or punishment), or the teacher can implicitly allow the boundary crossing by not responding.

Thus, important information about boundary control in the classroom would be contained in a record of the following kinds of boundary-related events: those resulting from a teacher instruction, from teacher permission, from a child's initiative but controlled by the teacher, and from a child's initiative and implicitly allowed by the teacher. A high-boundary classroom would be one in which many boundary crossings resulted from teacher instruction, or one in which few boundary crossings resulted from child initiative.

Another aspect of boundary control discussed previously, and included in the items of the Classroom Boundary Questionnaire, referred not so much to the boundary between each child and the surrounding social system but rather to the boundary around the task requirements. An observational estimate of the constraints imposed on children in doing their work would be contained in a record of the frequency with which the teacher gave specific directions about how to do the work, checked on the progress of work being done, or referred to time limitations.

The task of this study was to develop observational measures of teacher boundary control, to test their reliability and stability, and to investigate their relationship to teachers' scores on the Classroom Boundary Questionnaire.



### *Method*

Observations were done in 32 classrooms in the five schools of a single suburban elementary school system with a predominately white middle class and upper-middle class pupil population. There were 10 fourth grade, 11 fifth grade, and 11 sixth grade classrooms with a median class size of 21 (range 18-25). The distribution of classrooms by school was: 7 in school 1, then 9, 6, 5, 5 in schools 2 through 5. The eight men and 24 women teachers had a mean age of 34.5 ( $SD = 11.0$ ) and a mean number of years teaching experience of 8.8 ( $SD = 7.1$ ). The teachers volunteered to participate in the study after the investigator described the aims and procedures at faculty meetings at each of the schools.<sup>4</sup>

### *Observational Procedures*

As part of a larger study on teacher-pupil interactions in the classroom (Morrison, 1972), each classroom was observed for four half-hour periods between January and March by a single research assistant experienced in observing groups. She had been carefully trained in using the observational categories but was not aware of the specific hypotheses of the study. To establish reliability, the investigator observed along with the assistant for four sessions in each of three of the classrooms at the beginning of the study. Within constraints of scheduling, the observations were spaced over the 3-month period, and teachers were informed in advance of the schedule. Each classroom was observed during periods of instruction in several subject areas.

During each classroom visit, there were two four-minute periods during which *O* observed the classroom as a whole and noted each instance of four boundary-related behaviors: when children (a) left their seats, (b) talked with other children, (c) talked to the entire class, and (d) left the room. Frequency counts of these events are referred to as the classroom movement variables.

For each instance recorded, *O* also noted how the behavior was initiated: (a) by teacher instruction, (b) by teacher permission, or (c) by the child's own initiative. If the behavior was initiated by the child, the observer watched for the teacher's reaction and noted whether the teacher (a) tried to stop the behavior or (b) implicitly

<sup>4</sup> Out of a total of 45 teachers in the school system at the requisite grade levels, 35 actually volunteered; but three sixth grade teachers at one school were omitted from the sample because their team teaching methods differed dramatically from the instructional procedures of other teachers in the system.

allowed the behavior by not responding.<sup>5</sup> In most cases it was easy to record boundary crossing behavior with this system. It was more complicated when class projects required many children to be out of their seats. In such cases *O* first counted the number of children out of seat and then scanned the room looking for conversations among children.

From these tallies four measures of boundary control were derived for each teacher for each period of observation: (1) Teacher Instruction was the sum of teacher-initiated events; (2) Teacher Permission was the sum of events for which the teacher gave permission before they happened; (3) Teacher Control was the number of times the teacher reprimanded a child or told a child or the class to stop a behavior; (4) Child Initiative allowed was the total number of events initiated by children in the classroom that the teacher did not try to stop.

During the four-minute observation period, *O* also had the task of recording teacher comments that reflected attempts to control the boundary around the task. Three types of teacher statements were included: (1) specific directions about assigned work (statements in the imperative mode about work at hand); (2) directions about specific procedures (references to the proper or approved way of doing work); (3) references to the progress of work or to time limits. The variable called Task-Related Directions was the total number of statements in these three categories. This variable measured aspects of teacher control over time boundaries and behavior boundaries in the classroom.

The five variables described above, teacher instruction, teacher permission, teacher control, child initiative allowed, and task-related directions, are referred to as the teacher behavior variables.

### *Questionnaire Administration*

After all observations had been completed, the investigator administered a set of questionnaires in each classroom. These included the My Class Inventory (Anderson, 1971), with several items added to determine the children's perception of boundaries in the classroom (sample: "It is all right to get up and walk around in class"). A mean score on this scale was computed for each classroom. The classroom teacher completed her own set of questionnaires at the back of the room while the investigator administered questionnaires

<sup>5</sup> A list of examples of incidents that would be included under each category is available on request.

to the children. The teachers identified themselves by name on the questionnaires, having been assured of complete confidentiality. The children did not indicate their names.

### *Results*

#### *Reliability and Stability of Observational Measures*

Agreement between observers on the observational measures was assessed by the correlation coefficients between the investigator's and the observer's ratings over the 12 classroom sessions they observed together. They observed four sessions in each of three classrooms at the beginning of the study. The correlation coefficients for teacher behavior variables and classroom movement variables shown in Table 1 confirm that the investigator and the observer agreed on the categorization of the various behaviors.

The stability of each of the behavior categories over time was assessed by correlating the mean of the observer's ratings for the first and third observation periods with the mean of her ratings for the second and fourth observation periods. These coefficients for stability, corrected by the Spearman-Brown formula and shown in Table 1, are considerably lower than those for observer agreement. A repeated measures analysis of variance over the four observa-

TABLE 1  
*Observer Agreement and Stability for Observational Variables*

Variable	Observer Agreement <sup>a</sup>	Stability <sup>b</sup>
Teacher Behavior Variables	.99**	.61**
Teacher Instruction	.89**	.74**
Teacher Permission	.72**	.40*
Teacher Control	.65*	.55**
Child Initiative	.95**	.22
Task-related Directions		
Classroom Movement Variables	.91**	.39*
Children Out of Seat	.88**	.57**
Children Talking with Other Children	.97**	.51**
Children Talking Out	.96**	.22
Children Leaving Room		

<sup>a</sup> Observer agreement was assessed by the Pearson product-moment correlation between the investigator's and the observer's ratings over 12 class sessions that both observed.

<sup>b</sup> Stability of the observer's ratings was assessed by the correlation between the mean of the observer's first and third ratings of a classroom and the mean of her second and fourth ratings, over all 32 classrooms. This correlation was adjusted by the Spearman-Brown formula.

\*  $p < .05$ .

\*\*  $p < .01$ .

tional periods showed no main effects for time period. After these analyses, a mean for each variable over the four time periods was computed. These means were the data used in subsequent analyses.

### *Relationships among Observational Variables*

The only significant correlation among the observational measures of teacher behavior was between Child Initiative allowed and Teacher Control ( $r = .69, p < .01$ ). The relationships among the variables were explored by a factor analysis (Table 2) which revealed two factors with latent roots greater than 1.0 that together accounted for approximately 60% of the variance of the teacher behavior variables. A varimax rotation of the first two factors did not change the pattern of factor loadings found in the principal components solution. The first factor had three variables with high positive loadings: teacher permission, teacher control, and child initiative. The other two teacher behavior variables, teacher instruction and task-related directions, had high positive loadings on the second factor. The factor loadings, and especially the high positive correlation between the number of control attempts made by the teacher and the number of child-initiated events allowed by the teacher, suggest that an important underlying dimension in these classrooms was the amount of child activity.

### *Relationships with the Classroom Boundary Questionnaire*

Teachers' scores on the Classroom Boundary Questionnaire were related in the expected way to the amount of movement in their classrooms (Table 3). Score on CBQ, i.e., the amount of boundary control teachers said they preferred, was negatively correlated with the number of times children left their seats ( $r = -.40, p < .05$ ) and with the amount of child-child talk ( $r = -.44, p < .05$ ). The negative correlations with talking out and leaving the room were

TABLE 2  
*Principal Components Factor Analysis of Teacher Behavior Variables*

	1	2	Factors 3	4	5
Teacher Instruction	-.17	.51	-.84	.06	-.04
Teacher Permission	.61	-.33	-.17	.70	-.02
Teacher Control	.88	.19	.05	-.24	-.37
Child Initiative	.87	.01	-.17	-.28	.36
Task-related Directions	.14	.83	.44	.30	.10
Latent Root	1.94	1.09	0.96	0.73	0.28



TABLE 3

*Correlations Suggesting the Validity of the Classroom Boundary Questionnaire (CBQ)*  
*N = 32*

<i>Child Movement Variables</i>	<i>CBQ</i>
Out of Seat	-.40*
Talking with Classmates	-.44*
Talking Out	-.23
Leaving Room	-.29
<i>Teaching Behavior Variables</i>	
Teacher Instruction	-.13
Teacher Permission	-.31
Teacher Control	-.30
Child Initiative	-.48**
Task-related Directions	-.01

\*  $p < .01$ .

\*\*  $p < .05$ .

not significant but were in the expected direction. Further, the correlation between CBQ score and the questionnaire about boundary control that was administered to the children in each classroom was  $r = .59$  ( $p < .01$ ). That is, children in the classrooms reported experiencing the boundaries that the teachers said they maintained.

The correlations of the teacher behavior variables with CBQ are also shown in Table 3. CBQ score was negatively related to the amount of child-initiated boundary crossing that was allowed by the teacher ( $r = -.48$ ,  $p < .01$ ). A stepwise multiple regression analysis showed that the multiple correlation between CBQ and the teacher behavior variables increased only to  $R = .53$  when teacher permission and teacher instruction were included with child initiative as predictor variables.

### *Discussion*

This study suggests that the Classroom Boundary Questionnaire is a useful instrument for measuring teachers' preferences about boundary control in the classroom. The judges' consistent ratings in Study 1 suggested that the concept of boundary control was a clearly definable one; the factor analysis suggested that it was a unidimensional one. Study 2 showed that measures of teacher behavior derived from the concept of boundary control could be recorded reliably and were moderately stable across periods of observation. Teachers' scores on the CBQ were found to be in accord with the behavior observed in their classrooms: teachers who reported a preference for more control over the boundaries had classrooms in which there was in fact less movement. Most important,

there was a significant negative correlation between the degree of boundary control preferred by the teacher and the frequency of child-initiated boundary crossing events allowed.

Research relating to teacher control in classrooms began many years ago. Lewin and his colleagues studied the effects of authoritarian, democratic, and laissez-faire adult leadership on child behavior in small activity groups (Lewin, Lippitt, and White, 1939; White and Lippitt, 1968). Anderson used similar concepts of dominative vs. integrative leadership in his studies of classroom behavior (Anderson and Brewer, 1945; Anderson and Brewer, 1946; Anderson, Brewer, and Reed, 1946).

This early research was imaginative, and since then the concept of authoritarian or controlling behavior has been intuitively compelling to researchers. However, research applying these concepts to classroom interaction has not been fruitful (R. Anderson, 1959). The difficulties have been twofold: the research has not been based on psychological theory, and the concept of control has been confounded (Morrison, 1972; Smith and Hudgins, 1967; Wallen and Travers, 1963).

In most studies, controlling teachers have been defined as those who (a) set distinct limits that restrict the child's freedom of action and (b) respond to the breaking of limits in a punitive manner. The assumption has been that limit-setting and a tendency to be cold or punitive are highly correlated. Some data exist to suggest that this is not true (Christensen, 1960; Wright and Sherman, 1965). The confusion of these two aspects of teacher control has led to conflicting predictions about the effects of teacher control. In two recent studies one author predicted that higher teacher control would be associated with more stress and consequently less achievement in the classroom (Soar, 1967). Another author predicted that permissiveness would be negatively related to achievement (Christensen, 1960). The differences in expectation were related to different concepts of control. The differences in conception led to the use of measures of control or permissiveness that were so different that it is difficult to compare the results of the two studies.

The concept of boundary control is derived from a theory of social system functioning (Miller and Rice, 1967) and helps to clarify what is meant by teacher control in the classroom. As the leader of the classroom group, the teacher is responsible for creating conditions that allow the group to do its work. This means that the teacher must make decisions about how to control the interactions among the children in the group in such a way as to facili-

tate learning. This concept of control is not confounded. It clearly refers to the setting of limits, not to the degree of punitive behavior by the teacher. Thus, it is hoped that the measures of boundary control presented in this paper will be useful instruments in clarifying the study of teacher control in the classroom.

## REFERENCES

- Anderson, G. J. *The assessment of learning environments: A manual for the Learning Environment Inventory and the My Class Inventory*. Halifax, Nova Scotia: Atlantic Institute of Education, 1971. (mimeo)
- Anderson, H. H. and Brewer, H. M. Studies of teachers' classroom personalities I. Dominative and socially integrative behavior of kindergarten teachers. *Applied Psychology Monographs*, 1945, No. 6.
- Anderson, H. H. and Brewer, J. E. Studies of teachers' classroom personalities II. Effects of teachers' dominative and integrative contacts on children's behavior. *Applied Psychology Monographs*, 1946, No. 8.
- Anderson, H. H., Brewer, J. E., and Reed, M. F. Studies of teachers' classroom personalities III. Follow-up studies of the effects of dominative and integrative contacts on children's behavior. *Applied Psychology Monographs*, 1946, No. 11.
- Anderson, R. C. Learning in discussions: A resume of the authoritarian-democratic studies. *Harvard Educational Review*, 1959, 29, 201-215.
- Christensen, C. M. Relationships between pupil achievement, affect-need, teacher warmth, and teacher permissiveness. *Journal of Educational Psychology*, 1960, 51, 169-174.
- Getzels, J. W. and Thelen, H. A. The classroom group as a unique social system. In N. B. Henry (Ed.), *The dynamics of instructional groups*. 59th Yearbook of the National Society for the Study of Education, Part II. Chicago: University of Chicago Press, 1960.
- Jenkins, D. H. Characteristics and functions of leadership in instructional groups. In N. B. Henry (Ed.), *The dynamics of instructional groups*. 59th Yearbook of the National Society for the Study of Education, Part II. Chicago: University of Chicago Press, 1960.
- Jensen, G. The sociopsychological structure of the instructional group. In N. B. Henry (Ed.), *The dynamics of instructional groups*. 59th Yearbook of the National Society for the Study of Education, Part II. Chicago: University of Chicago Press, 1960.
- Ladd, E. T. The perplexities of the problem of keeping order. *Harvard Educational Review*, 1958, 28, 19-28. (a)
- Ladd, E. T. The problem of keeping order: Theoretical help from two new fields. *Harvard Educational Review*, 1958, 28, 136-149. (b)
- Lewin, K., Lippitt, R., and White, R. K. Patterns of aggressive

- behavior in experimentally created "social climates." *Journal of Social Psychology*, 1939, 10, 271-299.
- Miller, E. J. Technology, territory, and time: The internal differentiation of complex production systems. *Human Relations*, 1959, 12, 243-272.
- Miller, E. J. and Rice, A. K. *Systems of organization*. London: Tavistock Publications, 1967.
- Miller, J. C. Social process analysis. Mimeographed manuscript, Yale University, 1971.
- Morrison, T. L. Teacher control in school classrooms: A review of research. Unpublished manuscript, Yale University, 1971.
- Morrison, T. L. Teacher control of group boundaries in elementary school classrooms. Unpublished doctoral dissertation, Yale University, 1972.
- Roberts, J. I. *Scene of the battle: Group behavior in urban classrooms*. Garden City, New York: Doubleday Anchor, 1971.
- Schmuck, R. A. and Schmuck, P. A. *Group processes in the classroom*. Dubuque, Iowa: William C. Brown, 1971.
- Sheviakov, G. F. and Redl, F. *Discipline for today's children and youth*. Washington, D. C.: National Education Association, 1956.
- Smith, L. M. and Hudgins, B. B. *Educational psychology*. New York: Alfred A. Knopf, 1967.
- Soar, R. S. Pupil needs and teacher-pupil relationships: Experience needed for comprehending reading. In E. J. Amidon & J. B. Hough (Eds.), *Interaction analysis: Theory, research, and application*. Reading, Mass.: Addison-Wesley, 1967.
- Wallen, N. E. and Travers, R. M. W. Analysis and investigation of teaching methods. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963.
- White, R. K. and Lippitt, R. Leader behavior and member reaction in three "social climates." In D. Cartwright and A. Zander (Eds.), *Group dynamics: Research and theory*. New York: Harper and Row, 1968.
- Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.
- Wright, B. and Sherman, B. Love and mastery in the child's image of the teacher. *School Review*, 1965, 73, 89-101.
- Wright, B. and Tuska, S. A. From dream to life in the psychology of becoming a teacher. *School Review*, 1968, 76, 253-293.



## AN ASSESSMENT OF THE EFFECTIVENESS OF COMPLEX ALTERNATIVES IN MULTIPLE CHOICE ACHIEVEMENT TEST ITEMS

DANIEL J. MUELLER  
Indiana University

Discrimination indices and difficulty levels of multiple-choice achievement test items containing only substantive response alternatives were compared with items containing each of three complex alternative types: All of the above; None of the above; and combination complex alternatives (e.g., A and B; Either A or B; Two of the above, etc.). The same item statistics were compared, with items reclassified according to type of alternative keyed as the correct answer. Items containing combination complex alternatives were found to be most difficult, and items containing an "All of the above" alternative were found to be least difficult (especially when that alternative was keyed as the correct answer). Discrimination index was less affected by the inclusion or exclusion of complex alternatives than was difficulty level, but the highest discrimination occurred in items containing only substantive alternatives; the lowest in items in which "None of the above" was keyed as the correct answer. It was also found that all three types of complex alternatives functioned better as distractors than did substantive alternatives, with combination complex distractors receiving the highest rate of response.

WHILE there appears to be a high degree of agreement among test construction experts regarding the nature and importance of item writing principles, the empirical evidence supporting the validity of most of these principles is sketchy or nonexistent. This study will examine the effectiveness of three types of complex alternatives in multiple choice items: None of the above; All of the above; and "combination" alternatives (such as A and B, A and C, Either A or B, Two of the above, and the like).

Studies of the effectiveness of complex alternatives have had

mixed findings, Boynton (1950) concluded that use of the "None of these" alternative made items more difficult and better discriminating. Wesman and Bennett (1946) found the "None of these" alternative to be very effective in certain items, but not in all items. Rimland (1960) found no advantage in using the "Right answer not given" alternative. Williamson and Hopkins (1967) found that incorporating the "None of these" alternative in test items had no effect on test reliability or validity, but made items more difficult. And Hughes and Trimble (1965) concluded that complex alternatives can increase item difficulty (especially combination complex alternatives) but have little or no effect on item discrimination.

### *Method*

The present study compares discrimination indices and difficulty levels of items containing only substantive alternatives with those of items containing, respectively, each of the complex alternative types described above. A second series of comparisons follows the same pattern, but rather than grouping items according to the *presence* of a particular alternative in an item, items are classified according to the type of alternative which is keyed as the *correct answer*. Thus, discrimination indices and difficulty levels of all items in which a substantive alternative is the correct answer are compared with the same item statistics of all items in which "All of the above" is the correct answer, etc. Lastly, the usefulness of each of the three complex alternative types as distractors is examined.

Item statistics are from six unit examinations administered to students enrolled in the Indiana Approved Real Estate Salesmen's Course. Successful completion of this course is required by law before prospective real estate salesmen and saleswomen in Indiana can take the Real Estate Salesmen's License Examination. The course is offered three times a year at 28 locations throughout the state. Enrollment varies from around 300 in the Summer to between 700 and 1000 during the Fall and Spring semesters. The examinations utilized in this study are from four recent terms. Each examination contains one hundred items, more than half of which are multiple-choice. Each multiple-choice item has five alternatives. All examinations have KR-20 internal consistency coefficients between .84 and .86.

### *Results*

Table 1 shows mean difficulty level and mean discrimination index, by test and across tests, for items containing only substantive

TABLE 1  
*Mean Difficulty Levels and Discrimination Indices of Items Containing Only Substantive Alternatives and Items Containing Three Types of Complex Alternatives*

Examination	$N^a$	Substantive Alternatives Only		None of the Above			All of the Above			Combination Alternatives		
		$p^b$	$r^c$	$k^d$	$p$	$r$	$k$	$p$	$r$	$k$	$p$	$r$
Unit 1, Sp 72	1026	.72	.29	22	.59	.29	12	.58	.27	11	.62	.26
Unit 1, Sp 71	788	.85	.25	7	.72	.30	18	.76	.29	26	.66	.28
Unit 2, Sp 72	941	.83	.33	16	.77	.25	17	.83	.23	9	.66	.22
Unit 2, Fa 71	923	.81	.31	15	.74	.26	17	.84	.23	9	.61	.22
Unit 2, Su 71	294	.80	.30	15	.77	.25	15	.85	.26	12	.62	.20
Unit 2, Sp 71	670	.81	.32	16	.80	.26	15	.87	.27	12	.61	.21
Total		.79	.30	91	.74	.27	94	.78	.27	79	.64	.26

Note.—Items containing more than one type of complex alternative are classified in as many categories as they have complex alternatives.

<sup>a</sup>  $N$  = Number of students.

<sup>b</sup>  $p$  = Mean proportion of students getting items correct.

<sup>c</sup>  $r$  = Mean point biserial correlation of test items with total scores from their respective tests.

<sup>d</sup>  $k$  = Number of items.

alternatives, and for items containing, respectively, each of the three complex alternative types. On all but one test the highest mean discrimination index occurred in items containing only substantive alternatives. Items containing each of the three types of complex alternatives discriminated, on the average, about as well as one another. Differences in mean difficulty level were more extreme. Least difficult, on the average, were items containing only substantive alternatives ( $p = .79$ ) and items containing the "All of the above" alternative ( $p = .78$ ). Somewhat more difficult were items containing the "None of the above" alternative ( $p = .74$ ). By far the most difficult were items containing combination complex responses ( $p = .64$ ).

Item statistics in Table 2 are from items classified according to the type of response alternative which was keyed as the correct answer. While there was not a great deal of variance in discrimination indices, it appears that items in which a substantive response was the correct answer discriminated better, on the average, than did items in which any of the complex alternative types was the correct answer. Items in which "All of the above" was the correct answer were clearly easier, on the average, ( $p = .82$ ) than were items in which any other alternative type was the correct answer. Most difficult were items in which a combination complex alternative was the correct answer ( $p = .64$ ). Items with a substantive alternative as the correct answer and items with "None of the above" as the correct answer were intermediate in difficulty ( $p = .77$  and  $.76$  respectively).

Table 3 indicates the usefulness of each of the alternative types as distractors. On the average, each time a combination complex response was *not* the correct answer .13 of the students selected it, compared with .10 for "None of the above," .07 for "All of the above," and .05 for all substantive wrong alternatives.

### *Discussion and Conclusions*

Clearly, in this study, the inclusion or exclusion of complex alternatives had a marked effect on item difficulty, with items containing combination complex alternatives being the most difficult (whether or not the combination alternative was the correct answer), and items containing an "All of the above" alternative being the easiest (especially when that alternative was keyed as the correct answer). Discrimination indices were less affected by the inclusion or exclusion of complex alternatives. The highest mean discrimination index ( $r = .30$ ) occurred in items containing only substantive responses.



TABLE 2  
*Mean Difficulty Levels and Discrimination Indices of Items in Which, Respectively, Substantive Alternatives, and Each of the Three Complex Alternative Types is the Correct Answer*

Examination	N	Substantive Alternative			None of the Above			All of the Above			Combination Alternative		
		p	r	k	p	r	k	p	r	k	p	r	k
Unit 1, Sp 72	1026	.68	.29	35	.42	.25	1	.59	.28	4	.64	.26	13
Unit 1, Sp 71	788	.75	.28	23	.58	.25	1	.84	.30	14	.65	.30	9
Unit 2, Sp 72	941	.80	.29	32	.83	.29	2	.82	.19	6	—	—	0
Unit 2, Fa 71	923	.77	.29	32	.84	.21	2	.83	.21	6	—	—	0
Unit 2, Su 71	294	.80	.28	30	.83	.28	2	.85	.24	7	.62	.21	1
Unit 2, Sp 71	670	.80	.29	30	.82	.24	2	.88	.28	7	.61	.21	1
Total		.77	.29	182	.76	.25	10	.82	.26	44	.64	.27	24

TABLE 3  
*Mean Proportions of Students Responding to Substantive Alternatives and to Each of the Three Complex Alternative Types When They Are Wrong Answers*

Examination	N	Substantive Alternatives		None of the Above		All of the above		Combination Responses	
		p	k	p	k	p	k	p	k
Unit 1, Sp 72	1026	.08	49	.15	11	.06	7	.11	16
Unit 1, Sp 71	788	.05	46	.10	17	.10	12	.13	9
Unit 2, Sp 72	941	.05	39	.07	15	.01	3	.17	4
Unit 2, Fa 71	823	.05	40	.09	15	.04	3	.19	4
Unit 2, Su 71	294	.05	40	.10	13	.05	5	—	0
Unit 2, Sp 71	670	.04	40	.09	13	.07	5	—	0
Total		.05	254	.10	84	.07	35	.13	33

The lowest mean discrimination index ( $r = .25$ ) occurred in items in which "None of the above" was keyed as the correct answer.

Two qualifications are in order in generalizing from these findings. Examination of the proportionate use of complex alternatives as the correct answer relative to the inclusion of these alternatives (as right or wrong answers) in test items indicates that the "All of the above" alternative and the various forms of combination complex alternatives were overused as correct answers. In fact, "All of the above" was keyed as the correct answer 51% of the times it appeared in items, and combination alternatives were keyed as the correct answer 42% of the times they appeared in items. "None of the above" was seriously underused as a correct answer, being keyed as the correct alternative only 10 out of the 94 times it appeared in test items. This disproportionate use of complex alternatives as correct answers may have seriously affected item difficulty and discrimination. It is quite likely that if the "All of the above" alternative and the various forms of combination alternatives had been used more often as distractors, and if the "None of the above" alternative had been used more often as the correct answer, items utilizing these three forms of complex alternatives would have been more difficult and better discriminating.

The second qualification affects only the combination complex alternatives. This category contained several discrete kinds of alternatives. Consequently it is impossible to determine from this study the differential effectiveness of the various kinds of combination complex alternatives.

#### REFERENCES

- Boynton, M. Inclusion of "None of these" makes spelling items more difficult. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1950, 10, 431-432.
- Hughes, H. H. and Trimble, W. E. The use of complex alternatives in multiple choice items. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1965, 25, 117-126.
- Rimland, B. The effects of varying time limits and of using "right answer not given" in experimental forms of the U.S. Navy arithmetic test. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1960, 20, 533-538.
- Wesman, A. G. and Bennett, G. K. The use of "None of these" as an option in test construction. *Journal of Educational Psychology*, 1946, 37, 541-549.
- Williamson, M. L. and Hopkins, K. D. The use of "None-of-these" versus homogeneous alternatives on multiple-choice tests: experimental reliability and validity comparisons. *Journal of Educational Measurements*, 1967, 4, 53-58.





## THE ANALYSIS OF MULTIVARIATE GROUP DIFFERENCES

ALAN L. GROSS

The City University of New York

A computer program for evaluating the differences between  $k \geq 2$  groups on  $p \geq 1$  dependent variables is described. The statistical rationale for this program is based upon the Roy Union Intersection approach. A useful feature of the program is the computation of simultaneous confidence intervals for comparing the groups on each dependent variable and each discriminant function.

A common research problem encountered in the social sciences is that of identifying a set of  $p \geq 1$  variables that will discriminate among a set of  $k \geq 2$  groups. For example, a school counselor might wish to determine whether a set of biographical measures discriminates between student drop outs and non-drop outs. In a controlled experiment, one may ask whether  $k \geq 2$  experimental groups differ significantly from each other on  $p \geq 1$  dependent variables. The purpose of this paper is to describe the MANOVA Computer Program for ascertaining whether  $k$  groups differ significantly from one another on  $p$  dependent variables.

### *Rationale for the MANOVA Program*

A powerful technique for studying these multivariate discrimination problems is the Roy Union Intersection approach (Morrison, 1967). The null hypothesis to be tested is that the  $p$  by 1 mean vectors of  $k$  multivariate normal distributions are all equal.

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k; \mu_i' = [\mu_{i1}, \mu_{i2}, \dots, \mu_{ip}] \quad (1)$$

An equivalent statement of  $H_0$  is that for any linear combination of the  $p$  dependent variables, every comparison among the groups

has a value of zero. More specifically, letting  $D = \sum a_i x_i$  denote some arbitrary linear composite of the original dependent variables ( $x_1, x_2, \dots, x_p$ ), and  $c_1, c_2, \dots, c_k$  a set of comparison weights ( $\sum c_j = 0$ ), the null hypothesis can be stated as

$$H_0: \sum c_i \sum a_i \mu_{ji} = 0 \quad \text{for all } a_i \\ \text{for all } c_i; \sum c_i = 0 \quad (2)$$

The hypothesis of equal mean vectors (1) is true, if and only if statement (2) is true.

For a particular choice of the "a" weights and the "c" weights, a null hypothesis can be tested through a univariate  $F$  test. Every possible null hypothesis of the form  $\sum c_j \sum a_i \mu_{ji} = 0$  can be tested by choosing the values of the  $c_j$  and  $a_i$  that maximize the value of the  $F$  ratio and by then testing this maximized  $F$ . If the largest possible  $F$  value is not significant, then no other choice of the  $c_j$  and  $a_i$  can produce a significant  $F$ . In this case of a nonsignificant  $F$  ratio, there would be no significant differences between the groups. If the largest  $F$  is significant, one infers that there is at least one dimension (linear composite) that discriminates among the groups and that consequently the mean vectors are significantly different. This significant dimension is the first discriminant function.

When the overall null hypothesis is rejected, the Roy approach provides a Scheffé type post hoc analysis for examining the basis for rejecting  $H_0$ . This analysis furnishes tests of significance between every pair of groups on every individual dependent variable and also on every discriminant function. The experimentwise level of significance is controlled in this analysis, regardless of the number of tests performed.

### *The Program*

The computer program MANOVA is based upon the Roy approach. Program input consists of control cards specifying the number of variables, number of groups, group sizes, critical value, and variable format cards. Program output first consists of a test of the overall null hypothesis based upon the largest eigen root criterion. (The maximum  $F$  statistic is proportional to this root.) If the overall hypothesis is rejected, the program then provides a complete post hoc analysis of mean differences in terms of each individual dependent variable, and each of the discriminant functions. The standardized discriminant function weights and the group centroids in discriminant space are also computed.

The program was written in FORTRAN for an IBM 370 computer. A source copy of the MANOVA program as well as a program manual can be obtained by writing the author.

#### REFERENCE

Morrison, D. F. *Multivariate statistical methods*. New York: McGraw-Hill, 1967.





## COMPUTER PROGRAMS FOR ROBUST ANALYSES IN MULTIFACTOR ANALYSIS OF VARIANCE DESIGNS

PAUL A. GAMES

The Pennsylvania State University

Robust analyses of multifactor analysis of variance data may be secured even when assumptions are violated, by use of a set of five programs. ANOVR produces means, the summary table, group variances, the variance-covariance matrix on repeated measures, and tests most assumptions. FOLUP tests any set of pairwise contrasts of means, while COMCON tests any set of complex contrasts. Both latter programs permit solutions using heterogeneous group variances, or using a heterogeneous variance-covariance matrix from repeated measures, as well as conventional solutions using mean squares from the summary table.

ANOVR produces the Box-Geisser-Greenhouse index when there is just one repeated measure factor, while PVCVRL will produce corresponding indices when there are two or more such factors. BARTKI produces output that generates the robust and flexible Bartlett and Kendall test for homogeneity of variances. Intelligent use of the set of programs frees the user from most assumptions of AOV.

THIS paper describes a set of five programs providing robust techniques that work well when the assumptions of conventional analysis of variance (AOV) have been violated. The main program, ANOVR, includes relatively complete tests of assumptions and provides punched output that constitutes input data for three other programs. It can handle up to four between-subject factors and up to four within-subject factors and thus will permit a total of eight factors in a mixed design. References are made to sections in Winer (1971) that illustrate the designs and features described.

*Designs with Subjects Nested under Factors; Independent Group or Between-Subject Factorial Designs (Winer, 1971, Chapter 3, 5, 6)*

Proportional  $n$ 's are required. Homogeneity of cell variances is tested by Bartlett's test. Main effect means, simple-main means, or simple effect cell means may be tested using all  $+1, -1$  contrasts (all pairs) by means of program FOLUP. If the homogeneous variance hypothesis is retained in the equal  $n$  condition, Familywise Type I error rate may be controlled by employing the Newman-Keuls test, or the Tukey Wholly Significant Difference (WSD) (Games, 1971). If unequal  $n$ 's are present, a modified WSD may be used to control familywise error rate, or multiple  $t$ 's may be found to control Type I error per contrast.

The modified WSD consists of comparing a studentized range value,  $q(\alpha, K, df_w)/\sqrt{2}$ , to the conventional  $t$  statistic,  $(\bar{Y}_k - \bar{Y}_{k'})/(MS_w/n_k + MS_w/n_{k'})^{1/2}$ . If heterogeneous variances are indicated, individual cell variances are read, the  $t$  statistic is replaced by the Behrens-Fisher statistic, and  $df_w$  is replaced by the Welch  $df$  solution (Winer, 1971, p. 42). Howell and Games (1974) showed that these modifications of the WSD provide for a high degree of control of the familywise Type I error rate despite unequal variances, and a later study has extended this finding to the unequal  $n$  case. When the homogeneous variance assumption is true, the use of the Behrens-Fisher solution rather than the uniformly most powerful  $t$  test causes only a small loss of power for  $n$ 's greater than 10.

Program COMCON effects tests of complex contrasts by employing previously punched means and variances. The use of  $MS_w$  for the homogeneous variance case, or the  $s^2$  values of individual cells for the heterogeneous case is specified by the consumer. The heterogeneous variance solution is by the robust Welch generalized  $t$  (Welch, 1947). The Welch  $F'$  statistic, a robust alternative to the conventional  $F$ , is available as an option (Brown and Forsythe, 1974; Kohr and Games, 1974).

If the conventional equal  $n$ , homogeneous variance situation holds, the  $+1, -1$  or complex contrasts are so easily done on hand calculators that use of FOLUP or COMCON is unnecessary. However, these programs save a great deal of effort and time when applied to data containing unequal  $n$ 's and/or heterogeneous variances.

*Designs with Subjects Crossed with Factors; Repeated Measure Designs (Winer, 1971, Chapter 4)*

The ANOVR program differs from other general AOV programs in that it automatically computes the variance-covariance (VCV)

matrix on all repeated measures. Behavioral scientists should pay more attention to these matrices and to the correlation matrices that may be derived from them. Wiley, Schmidt, and Bramble (1973) have illustrated structural analyses testing models of such VCV matrices. The VCV are often a source of psychological interpretations as well as of information indicating that the conventional assumptions of  $F = MS_T/MS_{ST}$  are grossly violated.

Huynh and Feldt (1970) demonstrated that the mean square ratios are distributed as  $F$  with conventional  $df$  only if the population VCV matrix has properties that produce a Box-Geisser-Greenhouse index,  $\lambda$ , of 1.0. Box (1954) and Geisser and Greenhouse (1959) showed that the mean square ratio is approximately distributed as  $F(\lambda df_u, \lambda df_d)$ . The conservative solution is to employ a minimum value of  $\lambda$  so that the  $F(1, n - 1)$  is used. Collier, Baker, Mandeville, and Hayes (1967) demonstrated that control of Type I errors is maintained by estimating  $\lambda$  from the sample VCV matrix and multiplying the usual  $df$  by the estimated value. This solution has far greater power than does use of the conservative  $F(1, n - 1)$  distribution. With a single repeated measure, ANOVR computes  $\hat{\lambda}$  and reports the probability of a mean square ratio under all three solutions.

With two repeated measure factors at levels  $J$  and  $K$  respectively, ANOVR computes the  $JK$  by  $JK$  VCV matrix, and determines the probability of a mean square ratio using the conventional  $df$  and conservative  $df$  solutions. To obtain the marginal VCV matrices corresponding to the tests of main effects (e.g., see Winer, 1971, p. 552) this VCV matrix is submitted to program PVCVRL. PVCVRL also computes  $\hat{\lambda}_J$  and  $\hat{\lambda}_K$  as needed to adjust the  $df$  of the tests of main effects for the Collier, et al. (1967) solution. This process may be extended for additional repeated measure factors; a three factor example is given in the PVCVRL write-up. The correlation matrix from any VCV matrix is computed, if requested.

The test for compound symmetry of a VCV matrix (Winer, 1971, p. 596-599) is carried out on the original VCV matrix by ANOVR, while PVCVRL can conduct this test on any matrix read in or generated. Similarly the Machley test that Huynh and Feldt (1970) employ to test whether the population  $\lambda = 1.0$  is conducted by both ANOVR and PVCVRL. Since the compound symmetry case is a special case of  $\lambda = 1.0$ , the latter test is more pertinent.

Programs FOLUP or COMCON may be used with  $MS_E$  from the summary table for doing contrasts when  $\hat{\lambda} \approx 1.0$ , or with the punched VCV matrix for doing contrasts when  $\hat{\lambda}$  is substantially less than 1.0. Again solutions using  $MS_E$  are easily done by hand calculators, but

solutions employing the  $VCV$  matrix are more conveniently done on the computer. The latter solutions are completely general, but slightly conservatively biased if the conventional simple additive model assumptions are met.

For multifactor designs, the mean squares that may be needed for tests of means of one factor at a given level of another factor (Winer, 1971, p. 545) are automatically computed by ANOVR. These mean squares provide appropriate  $MS_B$  estimates when assumptions are met. Appropriate  $VCV$  matrices generated by ANOVR or PVCVRL may be employed in FOLUP and COMCON when the assumptions are not met.

*Mixed Designs; Subjects Nested under Some Factors but Crossed with Other Factors (Winer, 1971, Chapter 7)*

All of the features of the above two cases are included in mixed designs. In the simplest such design (Winer, 1971, p. 518), with  $a$  levels of the between factor, and  $b$  levels of the repeated measures factor, there would be  $a$  different  $b$  by  $b$   $VCV$  matrices, one for each independent group. These are computed by ANOVR and may be printed or punched as requested. From each such matrix, a  $MS_{SB}$  term is computed that would be the appropriate  $MS_B$  for testing  $MS_B$  computed on this group alone. Bartlett's test is used to test the homogeneity of these values (Winer, 1971, p. 522).

The  $a$  different  $VCV$  matrices are pooled and are tested for homogeneity by the Box technique (Winer, 1971, p. 595). The test for compound symmetry and the previously mentioned Machley test are conducted on the pooled  $VCV$  matrix by ANOVR. If an experimenter wishes to test the homogeneity of any subset of  $VCV$  matrices, this step may be undertaken by using PVCVRL. Similarly PVCVRL can provide a pooled matrix for any set of  $VCV$  or correlation matrices, and will test for homogeneity, compound symmetry, or  $\lambda = 1.0$ .

The value of  $MS_B$  is also computed separately for each of the  $a$  independent groups, and Bartlett's test is used to test for homogeneity (Winer, 1971, p. 521). If a significant  $AB$  interaction is observed, and if simple effect tests are desired, FOLUP or COMCON can be employed with the individual group variances or with the individual group  $VCV$ 's as appropriate. Thus it is possible to obtain robust tests despite violations of the conventional assumptions. Similarly, if the  $\hat{\lambda}$  computed on the pooled  $VCV$  matrix is substantially less than one, and if  $MS_B/MS_{B(A)}$  is significant, use of this  $VCV$



matrix in FOLUP or COMCON will provide a robust test on  $B$  main effect contrasts. With three and four factor mixed designs, the amount of output produced rises rapidly, particularly if the individual group VCV matrices are requested. However, intelligent use of these various outputs can furnish relatively powerful tests with substantial control of Type I errors, even when the assumptions are violated.

The fifth program is designed to provide robust tests of homogeneity of variance. Bartlett's test (Winer, 1971, p. 208) and other classical tests of variance are decidedly permissive when the populations sampled are leptokurtic rather than normal or platykurtic. The BARTKI program is designed to take data prepared for an analysis by ANOVR or BMD08V and to compute  $\log s^2$  values on subsamples of that data. These values may then be submitted to ANOVR to complete the Bartlett and Kendall test of homogeneity of variance (Winer, 1971, p. 219-220). In addition to being robust to non-normality, this test can be used to evaluate hypotheses about variances that correspond to main and interaction effect tests in multifactor designs. (See Games, Winkler and Probert [1972] and Gartside [1972] for further exposition of the Bartlett and Kendall test.) In general the Bartlett and Kendall test has lower power than the Bartlett test, but it protects against excessive risk of Type I errors. Thus the Bartlett test of ANOVR may be employed as a safe negative indicator. If it is not significant, the hypothesis of homogeneity of variance may be retained. However, if the hypothesis of homogeneity is rejected by Bartlett's test, it is desirable to confirm this conclusion by use of the Bartlett and Kendall test. The BARTKI program, which is written using the ANOVR conventions, produces punched output that may be directly "input" to ANOVR for final processing.

Write-ups and a tape copy of these programs may be secured by sending a small tape to the author. All but the rather small FOLUP program use dynamic storage allocation to minimize storage demands. ANOVR and PVCVRL are available in different sizes to fit different computer capacities. In each case, the smaller size versions delete various options and features to save space. On a small computer, the use of several small programs can accomplish what is possible in one run on a large version of the ANOVR program. The tape includes copies of the several programs, the data examples used in the write-ups, and additional data examples for testing.

## REFERENCES

- Box, G. P. E. Some theorems on quadratic forms applied in the study of analysis of variance problems: II. Effect of inequality of variance and correlation of errors in the two-way classification. *Annals of Mathematical Statistics*, 1954, 25, 484-498.
- Brown, M. B. and Forsythe, A. B. The small sample behavior of some statistics which test the equality of several means. *Technometrics*, 1974, 16, 129-132.
- Collier, R. O., Baker, F., Mandeville, G. K., and Hayes, T. Estimates of test size for several test procedures based on conventional variance ratios in the repeated measures design. *Psychometrika*, 1967, 32, 339-354.
- Games, P. A. Multiple comparisons of means. *American Educational Research Journal*, 1971, 8, 531-565.
- Games, P. A., Winkler, H. B., and Probert, D. A. Robust tests for homogeneity of variance. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1972, 34, 887-909.
- Gartside, P. S. A study of methods for comparing several variances. *Journal of the American Statistical Association*, 1972, 67, 342-346.
- Geisser, S. and Greenhouse, S. On methods in the analysis of profile data. *Psychometrika*, 1959, 24, 95-112.
- Howell, J. F. and Games, P. A. The effects of variance heterogeneity on simultaneous multiple comparison procedures with equal sample size. *British Journal of Mathematical and Statistical Psychology*, 1974, 27, 72-81.
- Huynh, H. and Feldt, L. S. Conditions under which mean square ratios in repeated measurements designs have exact  $F$ -distributions. *Journal of the American Statistical Association*, 1970, 65, 1582-1589.
- Kohr, R. L. and Games, P. A. Robustness of the analysis of variance, the Welch procedure, and a Box procedure to heterogeneous variances. *Journal of Experimental Education*, 1974, 43, 61-65.
- Welch, B. L. The generalization of students' problem when several different population variances are involved. *Biometrika*, 1947, 34, 28-35.
- Wiley, D. E., Schmidt, W. H., and Bramble, W. J. Studies of a class of covariance structure models. *Journal of the American Statistical Association*, 1973, 68, 317-323.
- Winer, B. J. *Statistical principles in experimental design*. (2nd ed.) New York: McGraw-Hill, 1971.

## A FORTRAN PROGRAM FOR SIMULATING EDUCATIONAL GROWTH WITH VARYING SCHOOL IMPACT<sup>1</sup>

JAMES M. RICHARDS, JR., NANCY KARWEIT, AND  
TRUMAN W. PREVATT

The Johns Hopkins University

To facilitate empirical investigations of longitudinal methodology, a computer procedure was developed to generate artificial data in which true growth scores are known. This procedure simulates the results of the Educational Testing Service Growth Study that pertain to the relationships among scores on a test of academic potential and scores on a test of educational attainment administered on four occasions. Similarly, the procedure simulates the results of Project TALENT that pertain to the correlation of community per capita income with average academic potential of students and with school resources. It is also assumed in the program that school resources determine school impact. The program user is allowed to specify the correlation between resources and impact and the extent to which schools vary in impact. Student scores and school means generated by this program are entered on separate output tapes.

A variety of statistical procedures has been proposed for overcoming the difficulties in assessing educational growth or psychological change (Cronbach and Furby, 1970). Because true growth scores are unknown in most longitudinal research, however, it has been difficult to compare the relative accuracy of these procedures. Accordingly, a computer procedure was developed to generate arti-

---

<sup>1</sup> Based on research supported by funds granted to the Center for Social Organization of Schools, The Johns Hopkins University, by the National Institute of Education. The opinions expressed do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

Copyright © 1975 by Frederic Kuder

ficial data in which these true growth scores are known. Such artificial data should facilitate empirical investigations of longitudinal methodology.

### *Underlying Model*

It is important that artificial data resemble real data as closely as possible to insure that the conclusions of methodological investigations will apply to the analysis of real data. Therefore, this program simulates selected aspects of the Project TALENT study of the American high school (Flanagan, Dailey, Shaycoft, Orr, and Goldberg, 1962) and of the Educational Testing Service (ETS) Growth Study (Hilton, Beaton, and Bower, 1971). Project TALENT provided intercorrelations among a variety of community, school, and student characteristics for a representative sample of high schools in the United States. In the ETS Growth Study, students were assessed initially with a measure of academic potential, the School and College Ability Tests (SCAT), and a measure of educational attainment, the Sequential Tests of Educational Progress (STEP). Subject to the usual attrition in longitudinal research, the educational attainment of these students was reassessed with STEP on three subsequent occasions.

Although in the typical computer procedure for producing correlated scores on two variables,  $A$  and  $B$ , both scores are generated at the same time, in this program the  $A$  scores usually are given and a corresponding set of  $B$  scores with the specified correlation  $\rho_{AB}$  between  $A$  and  $B$  is generated by the computer. Accordingly,  $B$  scores are created by the following equation:

$$B = \rho_{AB}A + Z \sqrt{1 - \rho_{AB}^2}$$

where both  $A$  and  $B$  have a mean of 0 and standard deviation of 1, and where  $Z$  is a random normal variate.

An important characteristic of many real data is that students are assigned to schools nonrandomly whereas in most statistical tests it is assumed that subjects are assigned to treatments randomly. This program permits the user to choose between random and non-random assignment. When students are assigned nonrandomly the program strives to reproduce the average correlation ( $\rho = .54$ ) between community per capita income and average academic potential of students estimated from the Project TALENT study. Specifically, a (pseudo) random normal variate is generated for each school and is treated as the per capita income of that school's home com-



munity. Then academic potential scores for the students at that school are created so that across schools the underlying correlation between income and average potential is .54 and so that the ratio of between schools variance to total variance on potential simulates the Project TALENT ratio.

The program also strives to reproduce the interrelations among academic potential and educational attainment on four occasions obtained in the ETS Growth Study. Several true score parameters were estimated in an earlier investigation (Richards, 1974) and used in this simulation. Specifically, each student's true score on initial educational attainment is generated through using the estimated true score correlation ( $\rho = .88$ ) between academic potential and initial attainment. Then a true gain score is produced for that student on the basis of a combination of several estimated parameters and is added to yield the true attainment score for that student on occasion 2. Similarly, gain scores are generated and added sequentially to yield true attainment scores on occasions 3 and 4. After the appropriate amount of random error is added to each score, the scores are transformed to the metric of the ETS Growth Study observed scores. This simulation procedure closely reproduces the ETS Growth Study results (Richards, 1974).

Finally, it is assumed in the program that community income determines school resources and that school resources in turn determine school impact. Specifically, a measure of resources is created for each school on the basis of the correlation ( $\rho = .25$ ) between community income and resources estimated from the Project TALENT results. There is little empirical basis, however, for estimating either the correlation between resources and impact or the extent to which schools vary in impact. Therefore, the program allows the user to specify both the correlation between resources and impact and the standard deviation of the impact variable. This standard deviation is specified in a form of a number between 0 and 1. When the standard deviation is set at .10, the average true growth scores are the same as those obtained in the ETS Growth Study for a simulated school with average impact, and are 10% higher than the ETS averages for a simulated school one standard deviation above the mean on impact. (The simulated data appear to meet the necessary assumptions for this manipulation even if the ETS data do not.) When the various scores for students at a given school are computed, the average growth scores are adjusted in accordance with school impact, and no other changes are made.

### Input

The input to this program consists of two control cards. The first control card provides the following program controls:

1. Columns 1-2. Number of tape drive for school tape.  
(See Output section of this paper.)
2. Columns 4-5. Number of tape drive for student tape.
3. Column 7. A 1 in this column suppresses writing of the school tape.
4. Column 9. A 1 in this column suppresses writing of the student tape.

The second control card includes the following parameters specified by the user:

1. Columns 1-3. *Number of schools*. The permitted range is 001-100.
2. Column 4. *Assignment of students to schools*. Students are assigned randomly if a 1 is punched in this column and non-randomly (i.e., in accordance with Project TALENT results) if a 2 is punched.
3. Columns 5-8. *Average number of students per school*. The permitted range, which is 0001 to 9999, should be set so that the product of the number of schools and the average number of students does not exceed 20,000.
4. Columns 9-10. *Standard deviation for number of students per school*. The permitted range, which is 00 to 99, should be set small enough relative to the average number of students to eliminate much chance of a negative number of students for any school (if such a negative number occurs, the program sets the number of students for that school at 1). When this parameter is 00, all schools will have the same number of students.
5. Columns 11-16. *Correlation between school resources and school impact*. The permitted range is +1.0000 to -1.0000. The decimal point is *not* punched in the card, so if this correlation was set at -.78, then -07800 would be punched in these columns.
6. Columns 17-18. *Standard deviation for school impact*. This parameter takes the form of a number between .00 and .99 (again the decimal is not punched). When this parameter is .00, schools do not differ with respect to impact.
7. Columns 19-25. *Random normal variate initialization*. This parameter must be a seven digit *odd* random number between 0000001 and 8388607 (in accordance with FORTRAN limita-

tions, this maximum is  $2^{28} - 1$ ). This number provides the starting point for the generation of random normal variates. It should be different for each independent set of simulated data (the identical sequence of normal variates will be generated each time the same seven digit number is used as the starting point).

### *Output*

This program provides the following three outputs:

1. *Student tape*. For each student, this tape includes a school ID number, a student ID number within that school, the three true gain scores, and both true and observed scores for academic potential and for educational attainment on four occasions.
2. *School tape*. For each school, this tape includes a school ID number, community per capita income, school resources, school impact, average academic potential as computed from per capita income, and the average true and observed scores for students at that school on academic potential and on educational attainment on four occasions.
3. *Correlation matrices*. In addition, a printed output provides observed score means, standard deviations, and intercorrelations for the academic potential variable and the various measures of educational attainment. The first of two matrices summarizes the relationships among scores for students at all schools combined and the second summarizes the relationships among school means. The second matrix is not computed when the number of schools is less than 25.

### *Limitations*

This program, which is written in FORTRAN IV, requires the equivalent of 6500 core locations on an IBM 7094 computer. Two tape drives for output are also required.

### *Availability*

For a copy of the source deck, write to Center for Social Organization of Schools, The Johns Hopkins University, Baltimore, Maryland 21218. Please enclose five dollars (\$5.00) to cover costs of reproduction, handling, and mailing. Purchase orders should be payable to The Johns Hopkins University.

### REFERENCES

- Cronbach, L. J. and Furby, L. How should we measure change—or should we? *Psychological Bulletin*, 1970, 74, 68-80.

- Flanagan, J. C., Dailey, J. T., Shaycroft, M. F., Orr, D. B., and Goldberg, I. *Studies of the American High School*. Project TALENT Monograph No. 2. Pittsburgh: University of Pittsburgh, 1962.
- Hilton, T. L., Beaton, A. E., and Bower, C. P. *Stability and instability in academic growth—A compilation of longitudinal data*. Princeton: Educational Testing Service, 1971.
- Richards, J. M., Jr. *A simulation study comparing procedures for assessing individual educational growth*. Baltimore: Center for Social Organization of Schools, The Johns Hopkins University, 1974.



# A COMPUTER PROGRAM TO TEST A REPEATED MEASURES HYPOTHESIS USING HOTELLING'S ONE-SAMPLE $T^2$ STATISTIC

PETER P. VITALIANO AND SILAS HALPERIN  
Syracuse University

The extent and direction of bias in an approximate test ( $T_A^2$ ) of a one-way repeated measures hypothesis was studied. Monte Carlo methods were used to simulate nine multinormal parent populations. Five thousand samples were drawn from each population and an  $F$  statistic (transformed from  $T_A^2$ ) was calculated for each sample. Nine sampling distributions, each containing 5,000  $F$ s, were then observed.

$T_A^2$  was shown to be very conservative—the proportions observed in the upper tails of the nine distributions were much smaller (in most cases one-half the size) than the proportions expected.

The exact  $T^2$  test was also presented along with the description of a program which computes this statistic. The program is recommended, over available package programs, because it requires minimum input.

THE usual way to test a repeated measures hypothesis:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_p \quad (1)$$

is to use a univariate  $F$  test. A sufficient assumption for this test to be exact is that the population covariance matrix  $\Sigma$  possess compound-symmetry (equal covariances and equal variances). An alternative way to test (1) is to use Hotelling's one-sample  $T^2$  test which does not require the compound-symmetry assumption.

The  $T^2$  statistic has the general form:

$$T^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu}_p)' S_p^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}_p), \quad (2)$$

where  $n$  is the sample size,  $\bar{\mathbf{x}}$  and  $\boldsymbol{\mu}_p$  are respective  $p$ -dimensional

vectors of sample treatment and hypothesized population means, and  $S_x^{-1}$  is the inverse of the sample  $(p \times p)$  covariance matrix.

Since the researcher typically wants to test whether the elements of  $\mathbf{y}_x$  are equal, and not necessarily equal to a set of specified values (i.e., (1)), the  $T^2$  in (2) is inappropriate. However, the vector variable in (2), namely  $\mathbf{x}$ , may be transformed to a new vector variable  $\mathbf{\bar{y}} = C\mathbf{x}$ , where  $C$  is a full rank  $(p - 1 \times p)$  matrix that is chosen so that the hypothesis  $H_0: \mathbf{y}_v = \mathbf{0}$  implies the hypothesis in (1).

Anderson (1958) and Morrison (1967) have shown that if the rows of  $C$  are linearly independent and if they are also the coefficients of contrasts, then  $T^2$  based on  $\mathbf{y}$ , will test the hypothesis in (1). This statistic is defined as:

$$T_v^2 = n(\mathbf{\bar{y}} - \mathbf{y}_v)' S_v^{-1} (\mathbf{\bar{y}} - \mathbf{y}_v), \quad (3)$$

where  $\mathbf{\bar{y}}$  and  $\mathbf{y}_v$  are respective  $p - 1$ -dimensional vectors of transformed sample treatment and hypothesized population means and  $S_v = CS_x C'$  is the sample transformed  $(p - 1 \times p - 1)$  covariance matrix. Given (3), it can be shown that

$$\frac{(n - p + 1)}{(n - 1)(p - 1)} T_v^2 = F(p - 1, n - p + 1).$$

Two standard textbooks (Winer, 1962; Kirk, 1968) present an inexact alternative to the univariate  $F$ . Although Winer states (in a footnote) that this statistic is only approximate, Kirk presents it as an exact statistic. This approach tests (1) by the approximate statistic:

$$\frac{n - p}{(n - 1)p} T_A^2 = F(p, n - p), \quad (4)$$

where

$$T_A^2 = n \mathbf{\bar{B}}' S_x^{-1} \mathbf{\bar{B}}. \quad (5)$$

One should observe two differences in (3) and (5). First,  $\mathbf{\bar{y}}'$  is a  $p - 1$ -dimensional row vector:

$$\mathbf{\bar{y}}' = (\bar{x}_1 - \bar{x}_p, \bar{x}_2 - \bar{x}_p, \dots, \bar{x}_{p-1} - \bar{x}_p),$$

whereas  $\mathbf{\bar{B}}'$  is a  $p$ -dimensional row vector:

$$\mathbf{\bar{B}}' = (\bar{x}_1 - \bar{G}, \bar{x}_2 - \bar{G}, \dots, \bar{x}_p - \bar{G}).$$

Second, because the grand mean,  $\bar{G}$ , is a linear combination of sample means,  $\bar{x}_p$ 's, the  $C$  matrix which produces the  $\mathbf{\bar{B}}'$  in (5) is not full

rank. Thus, if this  $C$  were used to transform  $S_x$ , the transformed covariance matrix would be singular. It is for this reason that (5) contains the untransformed covariance matrix,  $S_x$ .

In order to study the distributional properties of  $T_A^2$ , the authors empirically generated distributions of (4). Monte Carlo methods were used to create three multinormal populations for different numbers of treatments, namely  $p = 2, 3$ , and 5. From these populations, samples of different sizes ( $n$ ) were chosen so that the denominator degrees of freedom,  $n - p$ , would be equal to 10, 30, and 120 regardless of the size of  $p$ . Thus nine sampling schemes were formulated for each  $(p, n - p)$  scheme: 5,000 samples were drawn,  $T_A^2$  was calculated for each sample, and a sampling distribution of 5,000 observed  $F$ s was formed.

Table 1 presents the discrepancies between observed and nominal  $\alpha$  levels for the nine generated distributions. Because  $T_A^2$  is shown to be so conservative its use, by an unaware researcher, would provide a test with less power than might be anticipated.

Given these results, the following recommendations are offered for researchers who wish to use  $T^2$  to test (1). First, general programs available should be used with care: such programs as the Multivariate General Linear Hypothesis BMDX63 and Multivariate Analysis of Variance and Covariance BMDX69 (Dixon, 1968) require a reparameterization of design variables; the tabular results do not support the reparameterization scheme (in [5]) implied by Kirk (1968) and Winer (1962). It should be mentioned that Winer (1971), in his second edition, has provided a correct

TABLE 1

Discrepancies between Observed Proportions and Expected Proportions (Upper Critical Regions of .05 and .01) for the Nine Sampling Distributions									
$\alpha = \text{nominal expected critical region} = .05$									
Numerator Degrees of Freedom = $p$	2			3			5		
Denominator Degrees of Freedom = $n-p$	10	30	120	10	30	120	10	30	120
OBSERVED PROPORTION	.0186	.0126	.0146	.0252	.0214	.0190	.0266	.0286	.0242
$\alpha = \text{nominal expected critical region} = .01$									
Numerator Degrees of Freedom = $p$	2			3			5		
Denominator Degrees of Freedom = $n-p$	10	30	120	10	30	120	10	30	120
OBSERVED PROPORTION	.0036	.0022	.0022	.0042	.0024	.0034	.0056	.0052	.0046

transformation. Second, for less sophisticated users, such as those not familiar with programs that require transformations, a  $T^2$  program is offered which requires minimum input for use.

### *Input*

The data deck for each experiment consists of  $2 + n$  cards, where the first card contains the number of treatments and the number of subjects and the second card contains the variable format of the scores to be read in for each subject. Each of the next  $n$  cards (corresponding to  $n$  subjects) contains a score vector for each subject, where the format of each card's scores conforms to that specified on the second card. The program handles a maximum of 20 treatments.

### *Output*

The computer output includes: the number of treatments and subjects, a table of treatment means, standard deviations, and the corresponding covariance and correlation matrices, a vector of transformed means, the Hotelling  $T^2$  statistic, and the  $F$  ratio with its appropriate degrees of freedom.

### *Availability*

A listing of the program may be obtained from Peter P. Vitaliano, Department of Psychology, Syracuse University, Syracuse, New York 13210.

## REFERENCES

- Anderson, T. W. *An introduction to multivariate statistical analysis*. New York: Wiley, 1958.
- Dixon, W. J. (Ed.). *Biomedical computer programs*. Berkeley, California: University of California Press, 1968.
- Kirk, R. *Experimental design: Procedures for the behavioral sciences*. Monterey, California: Brooks-Cole, 1968.
- Morrison, D. F. *Multivariate statistical methods*. New York: McGraw-Hill, 1967.
- Winer, B. J. *Statistical principles in experimental design* (1st ed.). New York: McGraw-Hill, 1962.
- Winer, B. J. *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill, 1971.



## LPA2: A FORTRAN V COMPUTER PROGRAM FOR GREEN'S SOLUTION OF LATENT CLASS ANALYSIS APPLIED TO LATENT PROFILE ANALYSIS

BERTIL MARDBERG

University of Bergen, Norway

A computer program for solving latent profile analysis (Gibson, 1959) is presented. Green's (1951) solution for latent class analysis was used.  $N$  observations on  $p$  variables constitute the data matrix. The program estimates latent mean vectors and classifies the observation vectors to them. The result is a number of clusters of the observation vectors. A  $\chi^2$ -test of local independence is defined. Latent and observed discriminabilities are defined as the proportion of a variable's variance which is explained by between-variance. Discriminability constitutes the lower bound for estimated variable reliability. The program performs the analysis for a specified series of numbers of latent profiles and specified values on the diagonal of the product moment correlation matrix.

Latent profile analysis, LPA, for continuous variables was developed by Gibson (1959) as a generalization of latent class analysis, LCA (see Lazarsfeld, 1959). He suggested Green's (1951) solution of latent class analysis as an estimation method for latent profile analysis. LPA2 is a program for the solution of Green's estimation method applied to latent profile analysis. The program constitutes a system for the arrangement of observation vectors into homogeneous clusters from continuous variables.

Latent profile analysis is a submodel within the general system latent structure analysis, LSA (Lazarsfeld and Henry, 1968), which serves as a model for relations between manifest (observable) variables and latent (nonobservable) variables. Latent profile analysis can be derived from two assumptions: the first being that of

local independence and the second that of a discretely distributed latent variable.

The general model for LPA is:

$$r_{ijkl} \dots = \sum q_s z_{is} z_{js} z_{ks} z_{ls} \dots \quad (1)$$

where  $r_{ijkl} \dots$  are the product moments,  $i, j, k, l, \dots$  being indices for  $p$  variables;  $q_s$  is the proportion of subjects in latent profile  $s$ . There are  $m$  profiles. The  $z_{is}$  entries are the latent profile elements for variable  $i$  and profile  $s$  in standard scores. The right-hand side represents parameters in the model.

Green's solution uses information up to and including third order product moments, i.e., equations of the type:

$$r_{ijk} = \sum q_s z_{is} z_{js} z_{ks} \quad (2)$$

By using manifest information (the left-hand side) in (2), one can transform factors of product moments up to and including the second order into estimates of latent profiles ( $z_{is}$ ) and proportions ( $q_s$ ). The number of latent profiles can be less than or equal to  $p + 1$ . A solution of the model presupposes a given values for  $m$ , the number of latent profiles.

### *Description*

The program reads in an  $N \times p$  ( $N$  observations on  $p$  variables) data matrix. The variables are standardized to have means of zero and standard deviations of unity. Product moments up to and including the third order are computed. Through use of a given value for  $m$ , the matrix equations are solved—an outcome which give estimates of  $z_{is}$  and  $q_s$ . This step is done for different values on the diagonal to the linear product moment matrix ( $r_{ii}$ ). The program also permits an iterative process for stabilizing the estimates. Each observation vector is classified to the nearest latent profile, and the set of  $N_s$  observations which are classified to latent profile  $s$  constitutes a cluster. The mean vector for this cluster is an observed profile. The latent profiles which are not allocated any observation vectors are known as nonsense profiles. The observed profiles constitute the results of the clustering. The estimating and classifying procedures are undertaken for a series of values of  $m$ .

Each analysis of a series of the number of latent profiles,  $m$ , and of the values on the diagonal,  $r_{ii}$ , is evaluated for its fit to the model and for the predicted number of wrongly classified subjects. A  $\chi^2$ -test of local independence is defined after Anderson (1958, pp. 264-267). Discriminability is defined as

$$d^2(i) = \sum q_{is} z_{is}^2 \quad (\text{see Meredith, 1965}), \quad (3)$$

where  $d^2(i)$ , the proportion of a variable's variance which is explained by between-variance, constitutes the lower bound for estimated variable reliability.

Preliminary simulation studies indicate that in order to keep the percentage of wrongly classified observation vectors below about 15, the average observed variable discriminability should exceed .70.

### *Input*

Parameter card, format card, and observation vectors are needed. The number of variables, the range of expected number of latent profiles, and the range of diagonal values,  $r_u$ , are specified for the standard case.

### *Output*

The output consists of (1) overall means and standard deviations, as optional linear and triple correlation matrices and eigenvectors and eigenvalues of the linear correlation matrix; (2) for every given value of  $m$  and  $r_u$ :

Check of the numerical solution.

Test of local independence.

Distances between latent and observed profiles.

Estimated and observed profiles.

Observed profiles and within standard deviations in raw scores.

Latent and observed discriminabilities.

Mean discriminabilities.

Lists of the assigned subjects to the latent profiles (the clusters).

Optional: Within linear correlation matrices and eigenvectors and eigenvalues of these matrices;

(3) summary of the solution process. For every value of  $m$  and  $r_u$  the number of observed profiles and the mean discriminabilities are given.

### *Limitations*

The numbers of variables  $p$  and of profiles  $m$  are restricted as follows:  $p \leq 30$ ;  $m \leq 25$ . The program takes approximately 40K words of core.

### *Computer and Program Language*

The program is written in FORTRAN V for a UNIVAC 1110 in double precision.

*Availability*

A copy of the UNIVAC-version (for which the reader is asked to provide a tape), a copy of the source listing, and a write-up with test data may be obtained from Bertil Mårdberg, Department of Psychometrics, Institute of Psychology, University of Bergen, P.O. Box 25, N-5014 Bergen U. Norway.

## REFERENCES

- Anderson, T. W. *An introduction to multivariate statistical analysis*. New York: Wiley, 1958.
- Gibson, W. A. Three multivariate models: Factor analysis, latent structure analysis and latent profile analysis. *Psychometrika*, 1959, 24, 229-252.
- Green, B. F., Jr. A general solution for the latent class model of latent structure analysis. *Psychometrika*, 1951, 16, 151-166.
- Lazarsfeld, P. F. Latent structure analysis. In S. Koch (Ed.), *Psychology: A study of a science*. Vol. 3. New York: McGraw-Hill, 1959, Pp. 476-543.
- Lazarsfeld, P. F. and Henry, N. W. *Latent structure analysis*. Boston: Houghton Mifflin, 1968.
- Meredith, W. Some results based on a general stochastic model for mental tests. *Psychometrika*, 1965, 30, 419-440.



## A POPULATION SUBGROUP MULTIPLE COMPARISON COMPUTER PROGRAM BASED UPON CATEGORICAL DATA

BERNARD A. RAFACZ

Navy Personnel Research and Development Center<sup>1</sup>  
San Diego, California

Much of the information obtained on a particular population within psychological and educational research involves categorical data, usually in the form of nominal responses to items on a questionnaire. It is often of interest to compare various sets of subgroups of the population with respect to their response distributions to the items on the questionnaire. Procedures for performing multiple subgroup comparisons have been presented by Snee (1974) and Gabriel (1966). The question considered is whether or not a set of subgroups is homogeneous over the item response categories. Furthermore, if a set of subgroups is declared heterogeneous for an item, which subsets of that set may also be declared heterogeneous? It could be necessary to perform subgroup comparisons based upon a combination of some categories of the item. The computer program presented in this paper allows a researcher to determine which subgroups of the population are heterogeneous for some combination of categories of a questionnaire item.

THE purpose of the FORTRAN IV computer program discussed herein is to perform multiple comparisons of the subgroups of a population with respect to their response distributions on questionnaire items. The program was designed to be flexible enough to be able to perform such comparisons for any number, sequence, and combination of subgroups of a population over any reasonable number, sequence, and combination of categories for an item. Finally,

<sup>1</sup> The opinions expressed are those of the author and do not necessarily reflect those of the Navy Department.

Copyright © 1975 by Frederic Kuder

the output is designed to be readily understood by persons with little or no background in statistics.

### Approach

Let  $S$  be a set of  $s$  subgroups of a population and  $C$  a set of  $c$  categories of a questionnaire item. Further, let  $n_{ij}$  be the observed sample of individuals from the  $i$ th population ( $i = 1, \dots, s$ ) responding to the  $j$ th category ( $j = 1, \dots, c$ ). The hypothesis considered is whether or not the subgroups  $S$  are homogeneous over the categories  $C$ . The subgroups  $S$  are asserted to be heterogeneous over the categories  $C$  if the following condition is met:

$$2I(S, C) > \delta,$$

where  $\delta$  is the upper  $\alpha$  percentage point of the chi-square distribution with  $(s - 1)(c - 1)$  degrees of freedom.  $I(S, C)$  is the likelihood ratio statistic [see Mood (1959)] given by:

$$I(S, C) = \sum_i \sum_j n_{ij} \ln n_{ij} - \sum_i n_{i.} \ln n_{i.} - \sum_j n_{.j} \ln n_{.j} + n \ln n,$$

where

$$n_{i.} = \sum_j n_{ij}, n_{.j} = \sum_i n_{ij}, \text{ and } n = \sum_i \sum_j n_{ij}.$$

The program user selects a collection of sets of subgroups of the population and the number of categories for each item to be analyzed. Utilizing the aforementioned test statistic, each set of subgroups is, or is not, asserted to be heterogeneous over the categories under consideration. If the subgroups in a comparison set cannot be asserted to be heterogeneous, the analysis is complete for this set for those categories. That is, as Gabriel (1966) pointed out, if a combination of subgroups cannot be asserted to be heterogeneous, then no combination of subgroups formed from that set could be asserted to be heterogeneous. However, if a set of subgroups is asserted to be heterogeneous, the program searches among all possible pairwise combinations of the set of subgroups for additional heterogeneous subgroups over the same categories. This analysis is performed upon each set of subgroups the user provides to the program.

Subgroup comparisons based upon the combining or collapsing of subgroups and/or categories can be made within the same run,

or subsequent runs. Comparisons on any combination of subgroups or categories is possible, limited only by the program array sizes. Each item may be re-evaluated any number of times within the same run for any desired combination of categories.

### *Input*

Input to the program consists of a sequence of information which describes in some way the groups that are to be compared and the items to be analyzed. The information that the program requires includes: (1) a general description of the population being considered; (2) the alpha level (Type 1 error) at which the groups are to be compared; (3) the number of characters (ICONT) to be read from a data record; (4) the total number of subgroups (NG) for which questionnaire item tabulations will be found; (5) the number of questionnaire items (NQ); (6) the unit (NU) on which the data for the subgroups is to be found; (7) a variable format statement for the data; (8) a set of NQ cards (each card in the set describes for a questionnaire item the location of the item in the data record, the characters on which tabulations are to be made for the NG subgroups, the number of categories of the item over which comparisons are to be performed, and any re-editing of the original data characters for this item analysis); (9) a set of cards for each of the NQ questionnaire items (each set describes the questionnaire item and its response alternatives); (10) a set of NG cards that describe the groups being compared; and (11) the number of sets of subgroups (NCOMP) being compared and the subgroups that are to appear in each comparison set.

### *Output*

The output for each questionnaire item analysis includes: (1) a description of the item and its response alternatives; (2) the names of the subgroups being analyzed; (3) the frequency of response of the subgroups over the item categories to include itemization of illegal responses; (4) a unique symbol associated with each set of subgroups being compared (if heterogeneity is not asserted among the subgroups, only that symbol occurs; otherwise the symbol and plus "+" sign occurs); and (5) a summary of all pairwise comparisons among subgroups for which heterogeneity was asserted.

For the case of zero frequency response of a subgroup to some questionnaire item category, the test of homogeneity is not performed for those comparison sets in which the subgroup appears.

However, all pairwise combinations of subgroups in a comparison set for which that subgroup does not appear are tested for possible heterogeneity.

### *Limitations*

The program is presently limited to include: (1) alpha levels .01, .05, or .10; (2) maximum of 30 subgroups; (3) maximum of 50 questionnaire items with a maximum of 12 response alternatives to an item; and (4) a maximum of 10 sets of subgroups (each set of subgroups is to be compared simultaneously over some of the response alternatives). All of the variable size limitations can be altered by increasing the array sizes within the program.

There is no provision built into the program for a user to combine arbitrarily subgroups within one program execution. However, because initial runs are usually necessary to decide which subgroups are to be combined, the appropriate subgroup collapsing can be considered on subsequent runs.

### *Availability*

A copy of this paper, a listing of the FORTRAN IV source program, a documentation package for users, and a sample problem may be obtained by writing to Bernard A. Rafacz, Navy Personnel Research and Development Center, Code 310 BR, San Diego, California 92152.

### REFERENCES

- Gabriel, K. R. Simultaneous test procedures for multiple comparisons on categorical data. *Journal of the American Statistical Association*, 1966, 61, 1081-1096.
- Mood, A. M. *Introduction to the theory of statistics*. New York: Wiley, 1959.
- Snee, R. D. Graphical display of two-way contingency tables. *The American Statistician*, 1974, 28, 9-12.

## WARD AND HOOK REVISITED: A TWO-PART PROCEDURE FOR OVERCOMING A DEFICIENCY IN THE GROUPING OF PERSONS

HUBERT S. FEILD

Auburn University

LYLE F. SCHOENFELDT

University of Georgia

The Ward and Hook (1963) hierarchical grouping program is a frequently used method to cluster persons into groups. Because of a deficiency in the procedure, the groupings are somewhat less than optimal. In order to meet this deficiency, a two-part procedure was developed to be used in conjunction with the Ward and Hook program for a more optimal grouping of subjects. The first part of the procedure checks the assignments of the subjects to the groups and removes inappropriately classified subjects. The second part confirms the reassignments of the subjects to their groups. Specifics regarding the application of the two-part procedure are discussed.

THE grouping of subjects into subgroups is a common practice used by many investigators. Numerous techniques have been employed to place people into subgroups which are homogeneous with respect to a number of grouping dimensions. One of the most popular and frequently used is the hierarchical grouping procedure developed by Ward and Hook (1963).

### *The Ward and Hook Procedure*

In many situations, an investigator has collected a series of measures (tests or other observations) on a sample of  $N$  subjects and desires to know which subjects have similar profiles on the variables. The hierarchical grouping procedure is an iterative one, the objective of which is to cluster systematically the  $N$  profiles on



the basis of their similarity. This procedure makes no assumptions as to the number of groups in the sample but instead begins by considering each of the  $N$  profiles until all are in one group. At each stage of the grouping, all possible pairs of the groups' profiles are considered and the two most similar ones combined. Through each of the stages, the total number of groups is reduced by one while minimizing the increment in total within-group error. An inflection in the incremental error from a given stage to the succeeding stage indicates that the groups combined were dissimilar, and the pairings at the stage preceding the inflection become the solution.

### *A Serious Deficiency*

One deficiency in the grouping procedure is that once assigned to a group, the individual remains in that group. As additional subjects are assigned to the groups, the profile of the group will shift leaving the original group member(s) on the periphery of group membership. Thus, the assignment of individuals to subsets is usually less than optimal at the conclusion of the grouping (Ward, 1963). In order to correct this deficiency, a two-part procedure has been developed to evaluate the fit of each subject to his assigned subgroup.

The first part consists of an "affirmation" program<sup>1</sup> which compares the profile of each subject with the profile of every subgroup and either affirms membership in the assigned group or removes him from it. Removal could be for the reason that (a) the subject was a "misfit" and should be reclassified to another group; (b) the subject was an "isolate" and should not be classified to any of the groups; or (c) the subject was an "overlap" who fit more than one group. Adjusted group means are computed following each change in group membership, and the process is repeated until the number of changes is minimal and the subjects' group membership is affirmed.

The second step provides confirmation by capitalizing on the fact that one has "known" groups. Discriminant functions are formed from the variables, and the groups are located in the discriminant space. The subjects are treated as "new" individuals and are classified to the groups (Cooley and Lohnes, 1971). Overlaps, isolates, and misfits are identified as described previously, and discrepancies in results, by comparing the discriminant function results with those obtained previously in the variable space, are noted. The net result is the classification of the subjects to the

<sup>1</sup> The program was originally written by Mike Brodie.

groups with 100% hits in either the variable or discriminant spaces.

The application of the two-part procedure described above results in a "cleaner," more optimal assignment of subjects to subgroups. As such, the subgroups formed on the basis of these procedures are more homogeneous in terms of the grouping variables than are groups formed without using the two-part procedure.

### *Output*

The output for the affirmation program consists of  $D^2$  (a measure of profile similarity) values indicating how well individuals fit the groups to which they have been assigned. Through use of a series of decision rules, evaluations are made for each subject concerning his group membership, i.e., misfit, overlap, or isolate. The output for the classification program is a discriminant analysis which classifies the subjects to groups. Decisions regarding the classification are made according to the suggestions of Cooley and Lohnes (1971) and Rulon, Tiedeman, Tatsuoka, and Langmuir (1967). The objective of this step is to obtain a 100% correct classification of subjects to their groups.

### *Program Description and Limitations*

Both programs which are written in FORTRAN IV have been used on IBM 360/50, 65, and 370/158 computers. The system can handle 1,000 subjects with a maximum of 30 variables on each. The limit on the number of groups is 25.

### *Program Availability*

The affirmation and classification program listings along with the decision rules to be used in evaluating the program results can be obtained by writing Dr. Lyle F. Schoenfeldt, Measurement and Human Differences Program, Department of Psychology, University of Georgia, Athens, Georgia, 30602.

### REFERENCES

- Cooley, W. A. and Lohnes, P. R. *Multivariate data analysis*. New York: John Wiley and Sons, 1971.
- Rulon, P. J., Tiedeman, D. V., Tatsuoka, M. N., and Langmuir, C. R. *Multivariate statistics for personnel classification*. New York: John Wiley and Sons, 1967.
- Ward, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 1963, 58, 236-244.
- Ward, J. H. and Hook, M. E. Application of an hierarchical grouping procedure to a problem of grouping profiles. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1963, 23, 69-81.



## ECHO: A COMPUTER BASED TEST FOR THE MEASUREMENT OF INDIVIDUALISTIC, COOPERATIVE, DEFENSIVE, AND AGGRESSIVE MODES OF BEHAVIOR

DAVID J. KRUS

University of Southern California

KAREL R. BALCAR

Charles University

PATRICIA C. BLAND

University of Minnesota

A military-economic type of allocation game was used as a framework for a personality, computer based test. Test scores were derived from the magnitudes of allocations of monetary units into four regions. This decision-making behavior was interpreted in relation to its underlying personality traits through using the imago algorithm. Standardization of test scores and their interpretations, based on variable test norms, were automated.

SINCE its beginning, the field of objective personality testing has been dominated by paper and pencil tests. This type of testing strategy measures, in the majority of cases, only reports *about* behavior, either actual or imagined. Rapid developments in the area of computer technology, paralleled by the increasing availability of time-sharing computing, make feasible several new strategies, which attempt to measure a group of personality traits by analyzing the *actual* behavior of the participant in a computer-simulated game.

### *The Allocation Game*

ECHO evolved from an extension and computerization of Hornstein and Deutsch's (1967) allocation game. This original three-parameter experimental game required subjects to channel their

efforts into several kinds of products; in this particular instance into coloring paper squares. Individualistic and cooperative modes of behavior were represented by black and blue squares, respectively, whereas red squares represented weapons. The computerized version of ECHO required subjects to make decisions about the allocation of monetary units into four interest areas. The instructions, displayed on a cathode-ray tube time-sharing computer terminal, read as follows:

This game is a model of an international situation. You can make decisions about the allocation of nine units (each unit representing 10 billion dollars) into the following areas:

*Independent Enterprise:* The expected payoff is double the original investment.

*Cooperative Enterprise:* The expected payoff amounts to three times the joint investment.

*Defense:* Zero payoff, but every unit invested by you protects your state against twice as many units directed against you by your opponent.

*Offense:* The payoff which you gain from your opponent is five times greater than the original investment made by you.

The political situation is uncertain and your mutual relationships with the other state (ruled by a computer named Cyber) are unclear. You can expect cooperation as well as aggression. Please make your decisions by entering nine whole units, separated by commas. Good Luck.

An example of a typical game is shown as follows:

ENTERPRISE		ARMED FORCES	
INDEPENDENT COOPERATIVE	DEFENSE	OFFENSE	
		? 5, 3, 1, 0	
LAYOUT			
PATRICIA	5	3	2
CYBER	3	2	4
PAYOFF			
PATRICIA	10	6	0
CYBER	6	6	0
		TOTAL SCORE	
		16	78
		12	120

Subjects enter their choices in response to a question mark at the end of the third line of each game. The subject's and computer's allocations are then displayed under the heading "LAYOUT." It should be noted that the choices allocated to defense are doubled.



"PAYOFF" for both the player and the computer is displayed together with summary scores.

### *The Imago Algorithm*

The conversation of ECHO from a strategy-oriented game into a personality test was accomplished by the introduction of the imago algorithm, a modification of the "tit-for-tat" algorithm (Balcar, 1967; Rapoport, 1965). The rationale of the imago algorithm is similar to the reasoning behind projective tests: if there is nothing specific in the stimulus configuration to determine a particular type of response, then the response must be solely determined by some aspects of the inner organization of an individual. The imago algorithm goes one step further: if a [simulated] interpersonal exchange is determined solely by decisions of a single individual, then these decisions reflect some aspects of the individual's personality. The imago algorithm is thus designed along principles similar to those incorporated into the electronic feedback circuitry, designed to amplify signals. Amplified are those aspects of the inner organization of an individual which determine the game-related decisions.

The full course of the imago algorithm, as programmed into ECHO, consists of four distinct stages. During the first stage, the subject is presented with a neutral set of stimuli. This presentation is followed by the first play period, in which the computer's strategy mirrors the previous responses of the subject. The first stage is then repeated to enable the subject to reevaluate his strategy and to prevent him from discovering the principles comprising the imago algorithm. The fourth and last stage of the test is again characterized by the mimic feedback (i.e., as in the first play period) inherent in the imago algorithm.

### *Parameters of the Play*

There are eight parameters of the play, which can be adjusted by the assignment of different values to variables on lines 140 through 210 of Version 2.0 of the ECHO program. These eight parameters and their corresponding descriptions are as follows:

The number of units to be distributed is NUNITS. This parameter is presently preset to nine. Assignment of different values to this parameter would probably have only a minor effect on the game.

The length of the delay period (IDELAY) is preset at three. This parameter disguises the reciprocity of the game. Low values asso-

ciated with this parameter maximize the amplifying effect of the imago algorithm, whereas higher values assure credibility of the simulated interaction.

The parameter determining the length of the game (IHALF) is preset to eight. The whole test in its present form therefore consists of 16 separate plays.

The combination of IDELAY and IHALF parameters determines the spacing and length of the tests's four component substages. As presently programmed, ECHO consists of four play periods: first neutral stage (plays one through three), first play period (plays four through eight), second neutral stage (plays nine through eleven), and second play period (plays 12 through 16).

The payoff parameters were preset on the basis of pilot experiments to the values of 2-3-[0,2]-5 for independent enterprise, cooperation, defense, and offense (IPFF - ICFF - [IDFF, IDBON] - IOFF), respectively. These values were carefully balanced to assure as closely as possible an equal probability for the occurrence of each type of response.

### *Test Scores, Norms, and Interpretation*

Four primary scores are indicative of the intensities of behavior in the independence, cooperation, defensiveness, and aggression regions. Their magnitudes are determined by the mean amounts of money allocated by the subjects to these respective areas. Other scores are possible, pertaining to the allocation trends and overall success of the player.

Variable test norms are implemented and updated after each testing run. This feature provides for an immediate analysis of each particular game. Interpretation of the game can be suppressed or used for the discussion of the game with the participant. An example of an interpretation is shown as follows:

Patricia, your scores for independence, cooperation, defensiveness, and aggression are 3.50, 1.75, .81, and 2.94. I administered this personality test to 70 subjects and as compared with them . . . Apropos, do you like to compare yourself to others? [No]. O.K. but you will never know. Changed your mind? [Yes]. In terms of T scores (where mean = 50 and standard deviation = 10) your standing on these four scales is 52.31, 44.22, 48.70 and 52.77. In plain language, you are: autonomous, cooperative, protective, and combative. Your gains were average. Thank you for playing with me. Take care. Bye.

### *Test Utility and Applicability*

ECHO was developed as an attempt to improve the classic paper and pencil test format by means of a testing strategy which

measures actual behavior of both theoretical interest and practical significance. Pilot experiments showed moderate correlations between test scores and several factorial scales of Cattell's Sixteen Personality Factors Questionnaire (Cattell, Eber, and Tatsuoka, 1970), as well as the scales of Rosenzweig's Picture-Frustration Study (Rosenzweig, Fleming, and Clarke, 1947). These preliminary findings should be followed by test development and validation prior to making any conclusions in practical testing situations. Contingent upon standard development and validation procedures, ECHO has the potential of becoming a valuable addition to the behavioral researcher's testing inventory.

### *Availability*

For extended documentation write to Research and Development Center, 13 Pattee Hall, University of Minnesota, Minneapolis, Minnesota 55455. Please specify OP-30, 1974.

### REFERENCES

- Balcar, K. R. Hra jako model rozhodovani. Unpublished M.A. thesis. Prague: Charles University, 1967.
- Cattell, R. B., Eber, H. W., and Tatsuoka, M. M. *Handbook for the Sixteen personality factor questionnaire*. Champaign, Ill.: Institute for Personality and Ability Testing, 1970.
- Hornstein, V. A. and Deutsch, M. The tendencies to compete and to attack as a function of inspection, incentive, and available alternatives. Unpublished manuscript. Columbia University, 1967.
- Rapoport, A. Additional experimental findings on conflict and games. University of Michigan: Mental Health Institute Report No. 168, 1965.
- Rosenzweig, S., Fleming, E. E., and Clarke, E. J. Revised scoring manual for the Rosenzweig picture-frustration study. *Journal of Psychology*, 1947, 24, 165-208.



# A PROGRAM FOR COMPUTING RANK CORRELATIONS FROM ORDERED CONTINGENCY TABLES

LEWIS R. AIKEN  
Sacred Heart College

Formulas and a FORTRAN program for computing Kendall's tau as well as a generalized Spearman rho coefficient from ordered contingency tables are described. Relative advantages and disadvantages of tau and rho as measures of association are considered. The program can be used for analyzing ranked data from many subjects on two variables or for comparing the responses of one subject with the correct responses to an ordered multicategory item.

THE problem of what to do with tied or identical ranks when computing rank-order correlations is usually not dealt with satisfactorily in psychological and educational statistics books. Glass and Stanley (1970) discussed the problem of tied ranks in more detail than the author of many other books, but they referred the reader to Kendall (1970) for other developments. Kendall (1970) generalized the problem of tied ranks to ordered  $R \times C$  contingency tables, and provided two tau coefficients for this situation:

$$\tau_b = 2S/\sqrt{TU}, \text{ and} \quad (1)$$

$$\tau_c = 2S/\left[N^2\left(\frac{m-1}{m}\right)\right]. \quad (2)$$

In these formulas,

$$S = \sum_{j=1}^C \sum_{i=1}^{R-1} \left[ f_{ij} \left( \sum_{k=j+1}^C \sum_{g=i+1}^R f_{gk} - \sum_{k=1}^{j-1} \sum_{g=i+1}^R f_{gk} \right) \right],$$

$$T = \sum_{j=1}^C \sum_{i=1}^R f_{ij} \left( \sum_{j=1}^C \sum_{i=1}^R f_{ij} - 1 \right) - \sum_{i=1}^C \left[ \sum_{j=1}^R f_{ij} \left( \sum_{i=1}^R f_{ij} - 1 \right) \right],$$



$$U = \sum_{i=1}^C \sum_{j=1}^R f_{ij} \left( \sum_{i=1}^C \sum_{j=1}^R f_{ij} - 1 \right) - \sum_{i=1}^R \left[ \sum_{j=1}^C f_{ij} \left( \sum_{j=1}^C f_{ij} - 1 \right) \right],$$

$$N = \sum_{i=1}^C \sum_{j=1}^R f_{ij},$$

$m = (R + C - |R - C|)/2$ ;  $f_{ij}$  is the frequency (number of observations) in the  $ij$ th cell of the contingency table.

Generalizing Spearman's rank-order correlation to ordered contingency tables, the following formula was derived by the writer:

$$\rho = (D_{\max} + D_{\min} - 2D)/(D_{\max} - D_{\min}), \quad (3)$$

where

$$D = \sum_{i=1}^C \sum_{j=1}^R f_{ij}(i - j)^2,$$

$D_{\max}$  is the maximum possible value of  $D$  and  $D_{\min}$  the minimum possible value of  $D$  for the given table of data.  $D_{\max}$  and  $D_{\min}$  can be quickly computed by a simple procedure devised by the writer. Unlike  $\tau_b$  and  $\tau_c$ , regardless of the marginal totals of the contingency table, the range of  $\rho$  is always  $-1.00$  to  $+1.00$ .

When  $R = C = 2$ , formula 3 reduces to:

$$\rho = \frac{N - |f_{11} - f_{22}| + |f_{12} - f_{21}| - 2(f_{12} + f_{21})}{N - |f_{11} - f_{22}| - |f_{12} - f_{21}|}. \quad (4)$$

And in the special case where the row marginal totals are identical to the column marginal totals, formula 4 further simplifies to:

$$\rho = 1 - 2(f_{12} + f_{21})/(N - |f_{11} - f_{22}|). \quad (5)$$

The values of  $\tau_b$ ,  $\tau_c$ , and  $\rho$  are quite similar when the marginal frequency splits are not too extreme. Otherwise, the tau coefficients have an advantage in that their probability distribution is known (see Kendall, 1970). On the other hand,  $\rho$  has an advantage over  $\tau$  in that the range of the former coefficient is always  $-1.00$  to  $+1.00$ . In the computation of  $\rho$ , greater weight is also given to larger differences between  $i$  and  $j$ .

Since it was considered potentially useful to know all three coefficients for ordered contingency tables, a FORTRAN program was constructed.<sup>1</sup> The program was written to compute the contingency tables and rank intercorrelations among 25 or fewer vari-

<sup>1</sup> Copies of the computer program and directions for its use will be sent on request by writing to: Lewis R. Aiken, P.O. Box 8884, Guilford College, Greensboro, N. C. 27410.

ables, although more variables can be easily accommodated by changing the DIMENSION statement. The output of the program consists of the observed cell frequencies, the marginal totals, and the values of  $D_{\max}$ ,  $D_{\min}$ ,  $D$ ,  $\rho$ ,  $\tau_b$ , and  $\tau_c$  for each pair of variables. The data deck is headed by a single parameter card punched according to format statement 1. The number of variables and the number of subjects, respectively, are punched in the first two four-column fields of this card. These two numbers are followed in successive two-column fields of the same card by the number of ordered categories in each respective variable. The rank data of a given subject on all variables are punched in successive columns of one card according to format statement 2.

In addition to analyzing ranked data obtained from many subjects, the program can also serve as a means of comparing the rankings given by a single subject with the correct rankings of a large group of items. The correct category placements of the items may be represented by the rows of a contingency table, and the categories in which the examinee actually places the items by the columns of the table.

#### REFERENCES

- Glass, G. V. and Stanley, J. C. *Statistical methods in education and psychology*. Englewood Cliffs, N.J.: Prentice-Hall, 1970.  
Kendall, M. G. *Rank correlation methods* (4th ed.). London: Charles Griffin, 1970.



## COMPUTER PROGRAM FOR THE SELECTION AND COMPUTATION OF MEASURES OF RELATIONSHIP

ROBERT A. SMITH, MARK JAMES, AND WILLIAM B. MICHAEL  
University of Southern California

Relative to given scale properties of each of two paired variables, a computer program for the identification and computation of the following indices of relationship is provided: phi, Spearman rank order, Kendall's Tau, Pearson's product moment (involving two continuous variables), biserial, and point biserial.

This program provides a procedure to identify in terms of the level of measurement which of the several measures of correlation would be most nearly appropriate for data analysis. The identification procedures are based on the recommendations of Glass and Stanley (1970, pp. 156-181) and Fox (1969, p. 232). Once the identification procedure is completed the program then computes the selected statistics employing the *IBM Scientific Subroutine Package GH20-0205-4* (IBM, 1970).

Statistics which will be computed for each criteria are indicated in Table 1.

### *Procedure*

The program deck must have a control card with these instructions to implement the procedure:

Column 1. If the desired statistic is known it can be ordered by the specified code (Format II):

- 1 = Phi Coefficient
- 2 = Spearman Rank Order Correlation
- 3 = Kendall's Tau
- 4 = Product Moment Correlation
- 5 = Biserial Correlation

TABLE 1  
Computer Routines in Relation to Scale Properties of Correlated Variables

ROUTINES	CRITERIA						
	Ordinal Data	Interval Data	Both Variables Dichotomous	One Variable False Dichotomy	One Variable True Dichotomy	Both Variables Continuous	One Variable Continuous
Phi Coefficient			X				
Spearman Rank							
Order Correlation	X						X
Kendall's Tau	X						X
Product Moment							
Correlation		X					
Biserial Correlation		X				X	
Point-Biserial				X			X
Correlation		X			X		



6 = Point-Biserial Correlation

0 or blank = the routine will use the criteria specified in Columns 2 through 4 to determine the appropriate routine to use

Columns 2-4. These columns specify level of data measurement.

It is possible to specify multiple levels if desired. If the routine cannot ascertain which routine will be used, it will print out a table indicating the routines available and the criteria used for each routine. If more than one routine meets the specified criteria then both will be run (Format 3A1):

A = Ordinal data

B = Interval data

C = Both variables dichotomous

D = One variable is a false dichotomy

E = One variable is a true dichotomy

F = Both variables continuous

G = One variable continuous

H = Both variables have multiple categories

Column 5. For the biserial routines this column identifies which variable is to be treated as a dichotomy. A code of 1 will identify the first variable. The program will default to the second variable with this column blank (Format I1).

Columns 10-19. Identifies the dichotomizing value for the first variable if needed. If none is indicated, the program will use the median value (Format F10.0).

Columns 20-29. Identifies the dichotomizing value for the second variable if needed. If none is indicated, the program will use the median value (Format F10.0).

Data cards follow the identification card. All data are read casewise for two variables (Format 2F10.0).

## REFERENCES

- Fox, D. J. *The research process in education*. New York: Holt, Rinehart and Winston, 1969.
- Glass, G. V. and Stanley, J. C. *Statistical methods in education and psychology*. Englewood Cliffs, N. J.: Prentice-Hall, 1970.
- IBM Corporation. *IBM scientific subroutine package GH20-0205-4* (5th ed.). White Plains, N. Y.: IBM Corporation Technical Publications Department, August 1970.



## INTERVAL SCALING USING PAIR COMPARISONS OR PAIR COMPARISON TREATMENT OF COMPLETE RANKS UNDER CASE III ASSUMPTIONS

ROBERT J. WHERRY, SR. AND CHESTER A. SCHRIESHEIM  
The Ohio State University

Two of the better approaches to the scaling of stimuli are the methods of pair comparison and pair comparison treatment of complete ranks under the assumptions of Thurstone's Case III. This paper outlines a computer program with four data input options which scales up to 40 stimuli using these methods. The program will also compute an absolute scale zero point if the user employs the Horst method of balanced values (acceptance or rejection of pairs) in addition to the pair comparison procedure. Output of the program is detailed and includes, in addition to scale values and standard deviations, several matrices which allow Case IV or V scaling by hand with minimal effort.

Two of the better approaches to the scaling of stimuli are the methods of pair comparison and pair comparison treatment of complete ranks. These procedures are most often applied under the assumptions of Thurstone's Case V because of the additional labor involved in scaling under Case III. Also, while the Horst method of balanced values (acceptance or rejection of pairs) yields an absolute scale zero point (Guilford, 1936), this method is usually not employed because of the computational effort involved. Since Case III scaling is more accurate than that for Case V (it does not assume equal discriminial dispersions, but instead incorporates variability into the computation of scale values) and because the Horst method produces scales which approximate ratio measurement, a computer program which allows the application of these methods is of considerable value to the scale-builder.

The THURSCALE program is designed to scale up to 40 stimuli

Copyright © 1975 by Frederic Kuder

using one of four data input options: (1) complete unordered pairs, (2) complete or incomplete preordered pairs, (3) ordered pairs with Horst balanced values, and (4) complete rankings of stimuli. Only one control card is required for operation, in addition to the data to be scaled.

The program is written in FORTRAN IV for the IBM 360/75 and 370/165, but should be compatable with nearly any FORTRAN compiler. It includes extensive comment cards which provide card input specifications and which label and describe each major subroutine. Output includes: (1) the matrix of input proportions, (2) the matrix of ordered proportions, (3) the matrix of ordered  $z$ -scores (useful if Case IV or V scaling by hand is desired), (4) Case III scale values for each stimulus, (5) standard deviations for each stimulus, and (6) scale distances between adjacent stimuli.

A listing and writeup of THURSCALE along with sample input and output can be obtained by writing to Chester A. Schriesheim, Department of Psychology, The Ohio State University, 404-C West 17th Avenue, Columbus, Ohio 43210. Punched deck copies are also available at cost.

#### REFERENCE

- Guilford, J. P. *Psychometric methods*. New York: McGraw-Hill, 1936.

## A SUBROUTINE FOR COMPUTING ITEM EFFICIENCY AND ASSOCIATED PROBABILITIES

RICHARD J. HOFMANN

Miami University

A subroutine for use in any item analysis program is provided. This routine computes the efficiency of a test item, the exact probability of obtaining this efficiency, and the chance probability of obtaining a greater efficiency for an item, given its difficulty level, discrimination index, and sample size.

THE efficiency index is a value ranging from zero to unity. It is a measure of the functional discrimination of a test item given its observed difficulty level and discrimination index. The greater the magnitude of the index the more efficient are the observed discriminations. It is computed as a function of both item difficulty and item discrimination (Hofmann, in press).

This subroutine is intended to be used in conjunction with any test analysis program that generates difficulty and discrimination indices based upon some upper-lower group split. Specifically, it will calculate item efficiency, the exact probability of the observed efficiency, and the chance probability of obtaining a greater efficiency index, given a particular sample size and the associated discrimination and difficulty indices.

In theory the subroutine is a special application of the Fisher exact test (as discussed by Hofmann, in press). The actual computations, however, are made through the use of the gamma function ( $\Gamma$ ) and natural base  $e$  logarithms ( $\ln$ ). Assume a two way table of the following form.



	Right	Wrong	
Group 1 (Upper)	A	B	A + B
Group 2 (Lower)	C	D	C + D
	A + C	B + D	A + B + C + D

The exact probability of obtaining this particular table given the restriction of fixed marginals is determined as  $p$ , where

$$x = \ln \Gamma(A + B + 1) + \ln \Gamma(C + D + 1) \\ + \ln \Gamma(A + C + 1) + \ln \Gamma(B + D + 1) \\ - \ln \Gamma(A + B + C + D + 1); \quad (1)$$

$$y = \ln \Gamma(A + 1) \ln \Gamma(B + 1) + \ln \Gamma(C + 1) + \ln \Gamma(D + 1); \quad (2)$$

$$p = e^{1/(x-y)}.$$

Following Hofmann (in press) the exact probabilities of successively less independent tables are computed and summed, under the restriction of fixed marginals, to yield the probability of obtaining a greater efficiency index given the item difficulty.

### *Language and Capacity*

This subroutine, which is written in double precision FORTRAN IV, was developed through the IBM system/360 with a  $G$ -level compiler. It requires 2066 bytes of storage. Of some concern is the total sample size, the maximum value of which is fixed by the double precision log gamma function argument range. On the IBM system/360 this maximum argument is fixed at  $4.2937 \times 10^{78}$ . Unfortunately, this number will vary according to the computer used.

The computation algorithm is very efficient, as it required slightly less than 3.78 seconds to compute the efficiency indices and associated probabilities for all possible 148 combinations of difficulty and positive discrimination indices for a sample of 26 as well as to default and correct itself for all possible 148 combinations of difficulty and negative discrimination indices.

### *Parameters*

The subroutine is called initially to establish a table of gamma values. All subsequent callings for each item must have the diffi-

culty level, discrimination index, and sample size for the item. The subroutine returns unchanged the item difficulty, discrimination, sample size and in addition the item efficiency, the exact and cumulative probabilities of this efficiency index.

### *Availability*

Copies of this manuscript, a listing of the subroutine, and a special small illustrative program are available. The illustrative program generates sample data and output. These materials may be obtained by writing to Richard J. Hofmann, Department of Educational Psychology, Miami University, Oxford, Ohio 45056.

### REFERENCE

- Hofmann, R. J. The concept of efficiency in item analysis. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, in press.



## A COMPUTER PROGRAM TO GENERATE SAMPLE CORRELATION AND COVARIANCE MATRICES<sup>1</sup>

RICHARD G. MONTANELLI, JR.

University of Illinois at Urbana-Champaign

Given a population-covariance matrix, this program can generate any number of sample-covariance (or correlation) matrices, based on any sample sizes. If no population matrix is input, the program generates random correlation matrices.

RESEARCHERS interested in conducting sampling studies with multivariate techniques like factor analysis need an efficient computer program to generate sample-correlation and/or covariance matrices, based on any given number of observations ( $N$ ). Given a population-correlation or covariance matrix, this program can generate sample-correlation (or covariance, if the input was covariances) matrices. It can also generate random sample-correlation matrices, based on normally distributed random numbers. Uses for such matrices in factor analysis have been reported by Humphreys and Montanelli (1975) and Montanelli (1974).

### *Method*

The program is based on the method discussed by Odell and Feiveson (1966). This method has the advantage that only  $n(n+1)/2$  ( $n$  = the number of variables (rows) in the correlation matrix) random numbers need to be generated in most cases (except when  $N < n + 30$ ). This method also saves a considerable amount

<sup>1</sup> Computer time was paid for in part by the Office of Naval Research under contract N00014-67-A-0305-0012, Lloyd G. Humphreys, principal investigator, and in part by the Department of Computer Science of the University of Illinois at Urbana-Champaign.

Copyright © 1975 by Frederic Kuder

of work in computing the correlation matrix, especially for large  $N$ . Uniformly distributed pseudorandom numbers are generated by the multiplicative method, first suggested by Lehmer (1951) and discussed by Jansson (1966, p. 33). Pseudorandom normal deviates are provided by transformations given by Box and Muller (1958), and pseudorandom chi variates are computed by using an approximation for the inverse function for more than 30 degrees of freedom (Abramowitz and Stegun, 1966, p. 941) or the method of Box and Muller (1958) otherwise.

### *Limitations*

The program is written in FORTRAN IV for the IBM 360/75. The work is performed by a subroutine COVGEN which is dynamically allocated by specifying two constants in the small main program. Thus, the program, which is totally independent of  $N$ , can be trivially altered for any  $n$  (up to the amount of memory the computer has available). The program can generate any number of samples with various  $N$ s from one or more population-correlation matrices and/or any number of random-correlation matrices in one run.

### *Availability*

A copy of this article, the program, and additional documentation are available from the author at the Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801.

## REFERENCES

- Abramowitz, M. and Stegun, I. A. *Handbook of mathematical functions*. U. S. Department of Commerce, National Bureau of Standards Applied Mathematics Series, 1966.
- Box, G. E. P. and Muller, M. E. A note on the generation of random normal deviates. *Annals of Mathematical Statistics*, 1958, 29, 610-611.
- Humphreys, L. G. and Montanelli, R. G., Jr. An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research*, 1975, in press.
- Jansson, B. *Random number generators*. Stockholm: Victor Pettersons Bokindustri Aktiebolag, 1966.
- Lehmer, D. H. Mathematical methods in large-scale computing units. *Proceedings of a Second Symposium on Large-scale Digital Calculating Machinery*. Cambridge, Mass.: Harvard University Press, 1951.



- Montanelli, R. G., Jr. The goodness of fit of the maximum-likelihood estimation procedure in factor analysis. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1974, 34, 547-562.
- Odell, P. L. and Feiveson, A. H. A numerical procedure to generate a sample covariance matrix. *Journal of the American Statistical Association*, 1966, 61, 199-203.



## BOOK REVIEWS

MAX D. ENGELHART, Editor

LEWIS R. AIKEN, JR., Assistant Editor

With this issue, Dr. Dennis M. Roberts of the Department of Educational Psychology of the Pennsylvania State University is replaced by Dr. Lewis R. Aiken, Jr., of the Department of Psychology of Guilford College as Assistant Book Review Editor.

The Book Review section of EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT is not a book approval section. Readers may expect adverse comments. We discourage invidious comparisons with competing books. We prefer to publish reviews on educational and psychological measurement, statistical and other methods applicable to educational and psychological research, and on applications of computers.

In general, reviewers are assigned books to review by the review editors with attention to their abilities and interests. Persons wishing to volunteer a review should query the book review editor to obtain permission in order to avoid undue duplication.

Michael J. Apter and George Westby (Eds.). *The Computer in Psychology*. New York and London: Wiley, 1973. Pp. xvi + 309. \$14.95.

Oriented toward the needs and interests of the psychologist, this short volume is organized into two parts: the first consisting of five chapters concerned with the basic principles and techniques of digital computers; and the second, of five chapters dealing with applications of computers to five areas of substantive interest. In addition to the contributions of one of the editors (Apter) to the writing of one entire chapter and of co-authoring another, five other psychologists were involved in the completion of the eight other chapters. That at the time of its preparation all authors were members of the Department of Psychology at the University College Cardiff in the University of Wales may account for the essential continuity of the text and for the relatively nonfragmented treatment of the topics covered. The resulting uniform level of

reading ease from this cooperative venture would suggest that this well-organized book with its extensive bibliography would be an appropriate text for the upper division student or first-year graduate student in psychology, although the first five chapters would be of interest to most students in the behavioral and social sciences.

In Part I the first two chapters are concerned, respectively, with an introduction to computers and with an introduction to programming. Each of these two chapters (each co-authored by John A. Wilson and Geoffrey Barrett) is easy to understand, reasonably current, and quite practical in its approach. At a relatively elementary level these first two chapters cover much of the same material about use of computers for data processing as can be found in other general texts as indicated by the annotated references at the end of the second chapter (one of which is, however, later than 1969). It is not until reaching the third chapter, which is concerned with the use of computer language in experimental control, that specific psychological concerns are emphasized. In fact, this third chapter authored by Geoffrey Barrett affords the experimental psychologist helpful background information about technology for on-line control of experiments. Extending the emphasis on computer methodology in experimental psychology, Chapter 4 by Godfrey Harrison entitled "The Computer in Psychology Experiments" includes material on calculation, collation, generation of displays, synchronization of presentation of temporally related stimuli, and production of runs of drawings of appropriate stimuli. Computer applications to the study of serial learning are also discussed along with use of computers in closed loop studies. Practical concerns are cited, and a helpful example is detailed. Finally in Chapter 5 by Michael J. Apter, the relevance of the computer is modelling of behavior from the standpoint of the use of structural and functional models is considered. Particular attention is given to the role of the computer as an information processing system analogous to that found in activities of the human nervous system. Interrelationships among theories, models, organisms, and computers are described and illustrated, and the technique of computer simulation is critiqued.

Although applications of computer technology to experimentation and modelling of behavior are set forth in Part I of the book, five relatively specific areas of application are developed in Part II. In Chapter 6 John A. Wilson details the use of the computer in experiments on the psychology of perception in both visual and auditory domains, and in Chapter 7 Godfrey Harrison describes how the computer has been employed in the study of the psychology of language. Numerous illustrative examples are cited in both chapters. A relatively short discussion of the computer in the study of animal behavior appears in Chapter 8, "The Computer in Com-

parative Psychology," by Stuart J. Dimond. In Chapter 9, James O. Robinson explains how the computer can be used in clinical psychology. In addition to storing clinical information, the computer affords a means for the automation of psychological testing and of interviews as well as provides a vehicle for test interpretation, diagnosis, and potentially even therapy. How the computer may be employed in education and training is the subject of the tenth and concluding chapter by Michael J. Apter and Geoffrey Barrett. As the reader might surmise, the origin and development of computer-assisted instruction (CAI) as well as its techniques are a central focus of this chapter. In the estimation of the reviewers, this chapter is one of the best they have seen regarding CAI, and the critique of CAI is an exhaustive and penetrating one.

It would appear that the contributors to this volume have succeeded admirably in introducing psychology students as well as professional psychologists who have not had much experience with digital computers not only to their key principles, their language, and their actual and potential uses as tools of psychological inquiry but also to their applications to a number of specific substantive areas. In deemphasizing data processing and statistical analysis per se and in treating the computer as a heuristic tool for both psychological theorizing and empirical investigation, the writers have made an important and perhaps unique contribution not only to the teaching of psychology but also to updating of knowledge of computers by research psychologists who now may be encouraged to apply what they have learned from study of this volume to their own areas of investigative interest.

WILLIAM B. MICHAEL

*University of Southern California*

JOAN J. MICHAEL

*California State University, Long Beach*

J. W. Atkinson and J. O. Raynor (Eds.). *Motivation and Achievement*. Washington, D.C.: V. H. Winston & Sons, (distributed by Halsted Press of John Wiley & Sons) 1974. Pp. xi + 479. \$19.95.

There has always been a fascination with the possibility that achievement is not just a function of competence or good fortune, that somehow the *willingness* to pursue a goal is likewise important. This fascination was given concrete form by David McClelland, John Atkinson, and others in their early attempts to assess an achievement motive, using fantasy responses to pictorial stimuli. And perhaps there are no two names which are more closely associated with the work on achievement motivation than these. Interestingly enough, however, following an initial shared interest in motivational assessment, their work has, in fact, subsequently



proceeded along definitely distinguishable paths. The McClelland path has been characterized by a series of fascinating construct validity studies which culminated in a most ambitious consideration of the origin and creation of achieving persons who in turn create achieving institutions and societies. The Atkinson path is characterized by a concern for a construction of a theoretical model—but a theoretical model which emphasizes the interaction of *persons* with *situations* to create achievement. The research reported and reviewed in this volume is part of the Atkinson path. Indeed, the book is primarily composed of papers by Atkinson and his co-workers—many of them previously published elsewhere. As such, it is a handy compendium of the Atkinson path in its current state. And the current state of affairs is—or at least should be—of interest to all those concerned with the nature and assessment of achievement.

The program of research reflected in this book is impressive. Of particular interest are the new directions this research has taken since the publication of an earlier summary volume (Atkinson and Feather, 1966). These new directions include a closer look at sex differences in achievement, a clear recognition of the importance of long-term goals and immediate performance, and a viable concern with application. Last, but certainly not least, this volume reflects further specification of the model which has, incidentally, also culminated in the development of a computer simulation routine.

Impressive as the work reflected in this volume may be, it nevertheless prompts certain questions and criticisms. First, I continue to wonder about the applicability of this theory of achievement to diverse social and cultural groups (see Maehr and Sjogren, 1971; Maehr, in press). The definition and measurement of the achievement personality is most obviously appropriate when talking about the white middle-class male of the Western world. Atkinson and his colleagues have, of course, recognized some of the problems here and have suggested approaches to solving them. Thus, Matina Horner's work on sex differences in achievement (see Chapters 6 and 13) does provide at least one explanation for the previous failure of the Atkinson model to apply equally to males and females. Of special importance in ultimately understanding the actualization of achievement motivation in different cultures may be the new emphasis on the role of the perceived instrumentality of a task (see Chapters 7 through 10). But recognizing such improvements over the past, one may still question whether the thematic apperception and test anxiety measures commonly employed in this research are so thoroughly embedded in and limited to one culture as to be limited in their application to various sociocultural groups. Interestingly enough, this volume includes scarcely a reference to the problems of cultural diversity in motivational orientation.

Second, questions about measurement are inevitable, especially since this program of research has been repeatedly subjected to considerable criticism in this regard (see, for example, Entwisle, 1972). This volume contains an interesting defense of the test-retest reliability problem (see p. 8ff.) in the use of thematic measures of achievement motivation and the addition of new data improves the case for validity. However, Atkinson and his colleagues have still not provided the practitioner with anything like a workable device or scheme for assessing achievement motivation. Be this as it may, what will be of greatest interest to measurement specialists is not the work on motivational assessment *per se*. Rather, it is the criticism of ability and achievement testing from a motivational perspective that should prove especially provocative. The thrust of the argument here is that the results of achievement and ability tests depend rather basically on the motivational orientations of the person tested. Indeed, Atkinson suggests that the results on these tests can justifiably be given a motivational as well as, or instead of, the traditional aptitudinal interpretation. Motivational patterns that are elicited in ability testing situations may be quite different than those elicited in the course of achievement in the "real world." Although this possibility has often been recognized, satisfactory theorizing is seldom to be found. Especially in his ACT research institute paper (chapter 20), Atkinson has proposed a coherent and heuristic theoretical interpretation that deservedly challenges traditional interpretation of ability assessment.

Finally, one may note that as the volume is a reiteration, expansion, and clarification of earlier themes, it contains no expanded reference to or perspective on certain recent developments in achievement theory. Thus, for example, Weiner's recent work on the reinterpretation of achievement motivation in attribution theory terms is virtually ignored. Although that is predictable, given the intentions of the book in playing out the implications of a previously developed system, it is still disappointing—disappointing for several reasons. Attributional analyses of achievement behavior are very much a part of the achievement theory scene at the moment. Possibly the analysis of achievement motive in terms of predilections to attribute causes variously to one's ability or effort or to situational factors such as luck or task difficulty, could provide the framework for more convenient assessment of achievement orientations. Conveniently administered and readily scorable tests of achievement attribution have, of course, been developed. The marriage of attribution and achievement theory might thus eventuate in a welcome replacement of fantasy measures of achievement orientation. Then too, the attributional analysis has provided an innovative perspective in studying the developmental pattern of

achievement motivation—a topic of some concern to educators but ignored in this volume and seldom effectively treated anywhere.

All in all, this is a book primarily for scholars concerned with achievement theory and related assessment problems. It provides a convenient updating of an important program of research in the area. It is possible that the volume could be effectively employed as an adjunct text or as a source of readings for upper-level courses in such diverse areas as educational/psychological measurement, counseling and career development, and motivational theory. Having some experience in using it for a course in the last area, I have considerable confidence of its applicability in this respect.

### REFERENCES

- Atkinson, J. W. and Feather, N. T. (Eds.), *A theory of achievement motivation*. New York: Wiley, 1966.
- Entwisle, D. To dispell fantasies about fantasy-based measures of achievement. *Psychological Bulletin*, 1972, 77, 377-391.
- Maehr, M. L. and Sjogren, D. Atkinson's theory of achievement motivation: First step toward a theory of academic motivation? *Review of Educational Research*, 1971, 41, 143-161.
- Maehr, M. L. Culture and achievement motivation. *American Psychologist* (in press).

MARTIN L. MAEHR  
University of Illinois, Urbana

Clinton I. Chase. *Measurement for Educational Evaluation*. Reading, Mass.: Addison-Wesley, 1974. Pp. viii & 312. \$9.95.

This excellent book should promote student acquisition of skills, knowledge, attitudes and interest in measurements. These goals include efficient writing of classroom tests and the use of appropriate standardized tests useful in the evaluation of achievement and in obtaining data needed for placement and guidance. Instead of detailed description of specific tests there is emphasis on the characteristics of various types of tests—their applications and their limitations. Such concepts as behavioral objectives, cultural bias, criterion versus norm referenced testing, kinds of achievement and ability scores are simply, but adequately, explained. Similarly, lucid explanations are presented with reference to the different types of validity and reliability.

The categories and related behaviors of the *Taxonomy of Educational Objectives* produced by Benjamin Bloom and others, including this reviewer, are effectively summarized. This is followed by presentation of the two-dimension grid used in classifying behavioral and content objectives without noting that this device was originated by Ralph W. Tyler. (see his *Principles of Curriculum and Instruction*.)

The writing and analysis of different types of objective items and essay questions are reasonably well explained, but no attention is paid to the writing of exercises relevant to quoted materials as exemplified in the *Progressive Education Eight Year Study*, the *Cooperative Study of General Education*, nor in this reviewer's *Improving Classroom Testing*. Too much emphasis is given to the evaluation of knowledge of facts and too little to evaluation of critical thinking skills.

The measurement of general ability or intelligence and of special aptitudes is effectively explained in Chapters 7 and 8. The important achievement batteries are discussed in Chapter 9. Chapters 10 and 11 are instruments useful in assessing personality traits including interests and attitudes. Chapter 12 deals with characteristics of test administrators which influence test performance and with such factors as test anxiety, fatigue, and response bias. Chapter 13 discusses testing programs and the interpretation test results to parents. The appendices contain brief but lucid explanations of the computation of means, standard deviations, and coefficients of correlation. A useful list is given of names and addresses of ten important test publishers.

#### REFERENCES

- Bloom, B. S., Engelhart, M. D., Furst, E. J. Hill, W. H., and Krathwohl, D. R. *Taxonomy of educational objectives. Handbook 1, Cognitive domain*. New York: McKay 1956.
- Engelhart, M. D. *Improving classroom testing*. What Research Says to the Classroom Teacher No. 31. Washington, D.C.: National Education Association, 1964.
- Executive Committee. *Cooperative study in general education*. Washington, D.C. American Council on Education, 1947.
- Krathwohl, D. R., Bloom, B. S., and Masia, B. B. *Taxonomy of educational objectives. Handbook 2, Affective domain*. New York: McKay, 1964.
- Smith, E. R., Tyler, R. W., and Staff. *Appraising and recording student progress*. New York: Harper Brothers, 1942. (The Eight Year Study of the Progressive Education Association.)
- Tyler, R. W. *Basic principles of curriculum and instruction*. Chicago: University of Chicago Press, 1950.

MAX D. ENGELHART

Harry Frank, *Introduction to Probability and Statistics: Concepts and Principles*. New York, N. Y.: John Wiley & Sons, Inc., 1974. Pp. xvi + 431. \$12.95.

#### (FIRST REVIEW\*)

\* The first of the two reviews was assigned by the review editor, the second by the former assistant review editor.



This introductory textbook is intended for a first Statistics course in the biological or social sciences. As such, the contents are fairly predictable. The level of mathematics required is minimal. Calculus is not used; topics such as summation notation and the binomial theorem are reviewed as needed. The treatment is extremely intuitive, and is one of the more successful examples of this approach. The style is even, only becoming more difficult toward the end of the text. The examples chosen require no special knowledge from a subject discipline.

The first of the 15 chapters deals with the elements of probability theory. Topics include events, the sample space, the addition theorem, conditional probability, independent events, and the law of large numbers.

Chapter Two treats discrete random variables and their probability distributions. The approach taken is to partition the sample space into a collection of mutually exclusive, collectively exhaustive events. Values of the random variable are then computed or assigned for each event. Probabilities are then associated with the values of the random variable, rather than with the events themselves. The idea of a probability distribution is presented in this context.

The next chapter discusses elementary finite counting problems. Topics include the fundamental principle of counting, permutations, combinations, and tree diagrams. A clear and thorough, but intuitive derivation of the binomial distribution is given. The connection with the binomial theorem is included.

Chapter Four presents the distinction between a population and a sample, between a discrete and a continuous random variable, and between grouped and ungrouped data. The next chapter is a standard treatment of the mode, the median, and the mean. The treatment covers both grouped and ungrouped data, and both samples and populations. The author carefully distinguishes between  $\mu$  as a population parameter and  $\mu$  as an expectation. A short section, together with an appendix, discusses the algebra of expectations.

Chapter Six presents the sample variance and population variance. The approach follows that of Chapter Five. Chapter Seven shows the method for standardizing a random variable. This is then illustrated with a binomially distributed random variable.

The next chapter sets out the basic facts about the normal distribution. A limit argument is applied to the binomial distribution. The histograms used are clear enough, but the final statement of the argument on page 151 is misprinted. The method for reading the table of cumulative normal probabilities is given, as well as the normal approximation to the binomial.

Chapter Nine discusses sample covariance, the Pearson product moment correlation, and the coefficient of determination. The discussion is limited to the descriptive uses of these statistics.



Chapter Ten treats sampling distributions and the properties of estimators. For point estimators the author defines the meaning of an unbiased estimator, a consistent estimator, and an efficient estimator. The property of linearity is not mentioned. The author states "... all unbiased estimators are indeed consistent, ..." This is not true, as can be shown with a simple example. Again, the author tells us that, for fixed sample size,  $\bar{X}$  the sample mean is more efficient (less variance) than any other unbiased estimator. The proof is incomprehensible, and at a minimum fails to prove the result. The chapter states the Central Limit Theorem, without proof, and shows the confidence interval for the mean.

Chapter Eleven presents hypothesis testing in general. Topics include the null and alternate hypotheses, level of significance, types of errors, effect of sample size, and the power of a test. The discussion is thorough. In this chapter and later, cases are presented where, using reasonable data, the results are statistically significant but not scientifically significant. In particular, the sample statistic calls for rejection of the null hypothesis even though the same statistic is even less likely under the null hypothesis. The author suggests a solution which will please few. Terming this occurrence: overpowering, he tells us to decrease the sample size, even if it means taking a subsample from data already collected. The reviewer considers this philosophy unsound.

Chapter Twelve continues the discussion of hypothesis testing as it relates to tests about means. The author systematically treats most of the combinations of: normal versus non-normal population, one mean or two means, variance known or variance not known, and small sample or large sample. The nonparametric tests are not covered. The introduction of Student's  $t$ -distribution as the appropriate density function for the quotient of two random variables is exceptionally well done.

Chapter Thirteen discusses the two most common tests for variances: the test for a single variance based on the Chi-square and a test for two variances based on the  $F$ -ratio. Chapter Fourteen sets out the Chi-square test as applied to goodness-of-fit, and as applied to contingency tables.

The last chapter, Chapter Fifteen, is an introduction to the analysis of variance. The discussion is clearer than most in justifying the basis of the method. The reviewer feels some well-chosen graphs would help if they would identify which variance goes with which distribution.

The topics the author has chosen to include are well presented, with a minimum of loose ends. However, topics which have not been mentioned include regression analysis, rank correlation, partial correlation, tests for significance of correlation, maximum likelihood estimation, and stratified sampling. This may pose a problem

in choosing a text for a second course if this text is chosen for a first course. In addition, it may be desirable to augment the text with additional exercises, or with a workbook, since there are few student problems requiring algebraic or arithmetic drill.

THOMAS CHURCH  
*Governors State University*  
*Park Forest South, Illinois*

### (SECOND REVIEW)

The book by Harry Frank is one of several that have come out recently designed to provide students in the social and biological sciences with an introduction to statistical methods. What is unique about the text is that it is intended for use in an introductory course at the undergraduate level. The prerequisite skills that are required to handle the book are spelled out as two years of high school algebra and "some mathematical maturity."

The book is divided into three parts: probability theory, properties of distributions, and statistical inference. In the first part, the author presents the various notions of probability, introduces the concept of a random variable, and ends the section with a discussion of permutations and combinations, and the binomial distribution. The second part of the book deals with descriptive statistics, standardization, normal distribution and the concept of covariation. The notion of expected values and the algebra of expectation are introduced very early in this part. The third part is concerned with estimation, sampling distribution, hypothesis testing, tests of hypothesis on means, variances and entire distributions, and ends with an introduction to analysis of variance. The book is concluded with an appendix which contains discussions of the summation notation, the partitioning of sums of squares, and a more formal treatment of the algebra of expectations than that given in the main body of the text. An attractive feature of the appendix is that it includes the solutions to the problems in the text.

The material in the text is presented in a lock-step fashion, with later chapters depending heavily on the concepts developed in earlier chapters. In general, the author introduces new concepts intuitively either in terms of solving a problem, or with the presentation of a concrete example. The concepts and principles inherent in the problem or the example are then abstracted and formalized mathematically. This approach is particularly useful in teaching students who lack mathematical training.

The author's treatment of most of the topics is lucid and concise. An important feature of the book that distinguishes it from many other introductory books is that the author introduces the notions of random variables and expected values very early and once having

introduced these, exploits them to their fullest advantage. This should instill in the students some appreciation for the nature of statistics. The discussion of such difficult topics as unbiasedness, consistency, and efficiency, though non-mathematical in nature, is extremely clear.

The chapter on hypothesis testing, despite being a condensation of chapters fourteen and fifteen of Bailey (1971), contains clear presentations of the types of error, and power. The author discusses both simple and composite hypotheses, the distinction between which is not usually found in introductory texts. The discussion that follows on scientific versus statistical significance is clear and instructive. In concluding the section, the author recommends that "... the investigator always cast a simple alternative hypothesis. . . ." However, the argument leading to this recommendation stops just short of being convincing. In later chapters, the author explains how composite alternative hypotheses could be interpreted as "synthetic" simple alternative hypotheses in connection with inferences about the mean in one and two sample problems. This more than makes up for the deficiency in the earlier discussion. The only fault in this chapter is the author's failure to point out the use of confidence intervals in hypothesis testing.

There are several other flaws that mar the usefulness of the book. In the expression for the standardized normal distribution, the author repeatedly includes the standard deviation outside the exponent. In the application of the central limit theorem to the distribution of the sample mean, the author's attempt to show that the sample mean is expressible as a sum of random variables is rather confusing. The author distinguishes between the algebra of expectations and algebra of variances. However, whenever he has to obtain the variance of a linear combination of random variables, he refers to the algebra of expectations.

The text contains a brief discussion on experimental design, too brief to be of any real value. In his discussion on matching, the author says, "The experimenter has access to these variables, however, and can therefore balance their effects by matching rather than by randomization." He goes on to say, "... for each 19-year old man in the experimental groups, he would assign a 19-year old man to the control group." The literature in social sciences abound in the misuses of the principles of matching. The author, by making these ambiguous statements, does not improve matters any. We feel that the author would have made the book more useful by including a more detailed discussion on the fundamentals of experimental design and their use in social research. The author includes an interesting and lengthy discussion on the derivation of Student's *t*-distribution. Nevertheless, the derivation is not instructive, and moreover, the discussion is almost identical to that found in Bailey

(1971). Given this, the author should have referred the interested reader to the above text and devoted the space to more important topics.

The chapter on analysis of variance, though mathematically correct for the most part, is rather poor in terms of the presentation. The rationale for not testing the means pairwise is not given. The author attempts to derive the test statistic intuitively. The derivation is misleading and erroneous, a fact that would have become obvious had the author attempted to reconcile this statistic with that developed more formally later.

The author's writing style leaves a lot to be desired. His utter disregard for mathematical grammar and his persistent use of "therefore" instead of "then" in conditional sentences, are annoying. The author's illustrations of statistical concepts with examples drawn from his own experience, although amusing initially, become tedious quickly. Yet another fault is his choice of examples to illustrate the procedures. These primarily revolve around Neandertal skulls and sports cars. Beginning students often profit from examples drawn from their own disciplines.

Despite these faults, the author is quite successful in introducing statistics via the underlying principles and concepts. There are at least two other books, Hays (1973) and Bailey (1971), that introduce statistics in a similar fashion. The mathematical level of Bailey (1973) is higher than that of either Hays (1973) or Frank. These latter texts are fairly comparable with regard to the mathematical skills required. However, Hays (1973) is more extensive in terms of both the topics covered and the discussions supporting the mathematics within each topic. Bailey (1971) and Hays (1973) are intended for an advanced undergraduate or a beginning graduate statistics course. Frank's book is intended for an introductory undergraduate statistics course, but it is unclear whether the text is intended for beginning or advanced undergraduate. Despite the faults mentioned earlier and the availability of competing texts, we feel the book would be of value for an introductory statistics course for advanced undergraduates.

## REFERENCES

- Bailey, D. E. *Probability and statistics*. New York: John Wiley and Sons, Inc., 1971.
- Hays, W. L. *Statistics for social sciences*. New York: Holt, Rinehart and Winston, Inc., 1973.

JAMES ALGINA AND HARI SWAMINATHAN  
*University of Massachusetts*



Leonard M. Horowitz. *Elements of Statistics for Psychology and Education*. New York: McGraw-Hill, 1974. Pp. xv + 464. \$10.95.

This book is designed as a basic textbook for an introductory, applied statistics course. In the Preface, the author alludes to the fact that there are a multitude of introductory/intermediate statistics textbooks on the market today, most of which are decent, respectable books with each author professing to offer something special. As is the case with these other authors, Mr. Horowitz feels that, based upon his teaching experience, he has developed certain pedagogical notions which he decided to translate into book form. These pedagogical notions take the form of avoiding the cookbook approach by first describing the statistical procedure and then explaining, through illustrations and examples, the theoretical underpinnings.

At first glance, the reader is struck by the similarity between this book and the above-mentioned multitude of books in terms of topics covered. While the author does not suggest a time frame for utilizing the book, it is assumed that the topics can be covered in a one-semester or two-quarter course and therefore, a second glance is necessary to determine first of all if there is any new sequencing of topics and secondly, whether the presentation of the material merits adoption of this book in light of the availability of other options.

Chapter 1 is an attempt to whet the reader's appetite for the subsequent chapters. It is this reviewer's opinion that this attempt to synthesize "Philosophy of Science," "Basic Research Methods," and the "Lying with Statistics" falls rather short of the intended goal. It may be viewed as a chapter that a professor might assign for students to "read casually."

As the author points out, there are two schools of thought regarding the inclusion of "old-fashioned" topics; i.e., frequency distributions and elementary descriptive statistics, in a textbook of this type. He adheres to his philosophy of not belaboring these simple topics in Chapter 2 and does a commendable job of illustrating the formula for percentiles. However, in Chapter 3, he belabors extensively. The concepts of Central Tendency and Variation, with the accompanying rules of summation, are again simple topics which are familiar to most students who would be using this text. It is this reviewer's opinion that while it is important for students to have "a feeling for" a frequency distribution in preparation for theoretical distributions, it is not necessary, and often more confusing than it is worth, to have students compute means, medians, and variances from frequency distributions. With modern com-



puters and packaged programs, these laborious and out-dated computational formulas should be forgotten.

Chapter 4 finishes the presentation of the basic statistical concepts with the normal curve and  $z$  scores. While the discussion of this first portion of the chapter is quite well done, two errors in content structure become apparent. First, the author suggests that the section on normalizing distributions can be omitted for a shorter course. However, this is the section in the book that deals explicitly with the scales of measurement, considered by many as an extremely important concept in applied statistics. Secondly, correlation is treated as a "related topic." Again, correlation is more than a related topic and deserves more than the eight pages which Horowitz has devoted to it. Additionally, there is no mention of other correlation coefficients, and the naive reader may be led to believe that the Pearson Product-Moment as defined in this book is "correlation." Chapter 12 refers only briefly again to correlation; however, it does have an excellent discussion of the relationship between linear correlation and linear regression. This section is especially meaningful in light of the growth of the Multiple Regression Approach in behavioral research.

While slighting measurement scales and correlation, the author deals exhaustively with probability and binomial distribution in the traditional set theory approach. At times, the reader gets bogged down with all the symbolism, and this may impede effective understanding of the concept of probability and how it relates to hypothesis testing and interval estimation. The soundness of the pedagogy in this exhaustive presentation may be questioned for this type of book.

The author is at his pedagogical best, however, as he begins the discussions of Statistical Theory and Hypothesis Testing. At this point, he begins to tie together all of the preceding concepts. The approach is quite logical with the author introducing new and sometimes quite difficult concepts for the most part in a very direct and readable manner. A weakness in this approach is the return to philosophy of science with the Latin terminology (p. 164) to explain the purpose of hypothesis testing and the associated errors inherent in this procedure. The student may have difficulty in learning and understanding the full meaning of rejecting or failing to reject the null hypothesis.

The discussion of the sampling distributions of the mean is very illustrative, and the direct tie to hypothesis testing and interval estimation is logically structured. This is a potential strength of the book in this reviewer's opinion due to the fact that many of the available books deal with these two topics separately and often do not tie them together. However, there is a subtle yet important flaw in the discussion of the confidence interval for the mean (p.

189-90). The author discusses the confidence interval in terms of the sample mean  $\bar{X}$  falling between the two extreme values; i.e., the confidence limits. He states that the 95% confidence interval implies that "the probability is .95 that the sample mean lies within 1.96 standard errors of the population mean," (p. 189) rather than that the probability is .95 that the interval  $\bar{X} \pm 1.96s_e$  spans the population mean. This is a critical error in reasoning and seriously effects the understanding of the relationship between hypothesis testing and interval estimation.

In the Analysis of Variance chapter, the author basically returns to the traditional cookbook approach to teaching ANOVA. While he does attempt to explain the meaning of the mean squares, sum of squares and degrees of freedom, the reader with a limited analytical mind and without a firm grasp on algebraic manipulation could easily get frustrated with all of the symbolism. Additionally, only one-way ANOVA is presented with an optimal section of a repeated measures design. In view of the rather exhaustive discussion of other topics earlier and later in the book, a more extensive discussion of higher order ANOVA and the various models (fixed, random, mixed) would seem to have been consistent with other sections of the book while also better explaining the method of ANOVA. This presentation was considered good at best.

When the author get to the chapters on non-parametric statistics, the introductory discussion is weakened due to the lack of a good discussion regarding the scales of measurement. While there is a continuing debate concerning the power of non-parametric versus parametric statistics for data measured on less than an interval scale, the scale of measurement is still regarded as one assumption underlying the use of parametric statistics, which the author fails to mention. The description of the various  $\chi^2$  distributions is well written but possibly more than is needed in a book of this type. The reader not totally familiar with ogive percentages and the concept of Expected Values may have trouble plowing through all of the explanations.

In summary, the author attempted to show the need for yet another elementary/intermediate statistics book by presenting the material in a more pedagogical way. While he generally succeeds, there are the obvious strengths as well as the glowing weaknesses. The strengths are primarily in the writer's logical presentation of the material. While the relevance of the exhaustive discussion of probability and the binomial and  $\chi^2$  distributions may be questioned, the material is well presented. Also well presented were the sections on the normal distribution, the t-distributions, the  $\beta$  error and the power of a statistical test. Major weaknesses include out-dated discussion of computations from frequency distributions, limited discussion of the concept and methodologies of correlation, limited

discussion of the scales of measurement, the critical flaw in the discussion of the confidence interval, and only a good presentation of ANOVA.

DENNIS E. HINKLE

*Virginia Polytechnic Institute and State University*

Arthur R. Jensen. *Genetics and Education*. New York: Harper & Row, 1972. Pp. vi + 378. \$10.00.

In this collection of previously published articles, Jensen has brought together and updated those most pertinent to the controversy which has become synonymous with his name. The cornerstone in the collection is the 1969 *Harvard Educational Review* (HER) paper which ignited the debate on the origin and nature of racial and social class differences in IQ, "How much can we boost IQ and scholastic achievement?"

As might be expected from such an emotionally charged issue, the debate has generated more heat than light, but as the impartial reader of these papers can see, the fault for this lies far more with his intemperate critics than with Jensen. This observation is strengthened by a most revealing preface, written especially for this book, in which Jensen chronicles professional and other reactions to the HER paper. It should give pause to those who feel that free speech and academic freedom are totally secure and impregnable.

The more serious criticism to which Jensen's claims have been justifiably subjected has concentrated on two major areas: the accuracy and completeness of the genetic arguments employed; and the nature and meaning of the measurement process, i.e., the IQ.

Two key points in the genetic argument are most frequently cited as weak links in the chain of Jensen's argument. The first has to do with the relationship between *within* group and *between* group heritability. In this edition the use of footnotes which cite DeFries' formula for the exact theoretical relationship helps to clarify the issue somewhat. It is difficult for non-geneticists to understand the arguments completely or to know the appropriate analogies to other kinds of data which might permit an estimate of the likely relationship, but two inferences can be drawn. First, a high *within* group heritability increases the a priori probability of a high *between* group heritability. Second, the actual figure is an empirical question which only the collection of relevant evidence can decide. Reasonable critics have suggested that such evidence could well fail to support Jensen's "not unreasonable" conclusion that genetic factors are implicated in racial and social class differences in IQ scores. But it should be noted that Jensen points out both of these



inferences in his book, and welcomes the research which could shed light on this dark corner.

The second area of criticism regarding the genetic argument is related to questions about the measuring instruments, and is framed by the question "When is a difference a deficit?" Some of Jensen's critics have argued that simply citing the intellectual component of middle class life as the ideal does not make it so, and the subsequent discovery of a mean score on IQ tests for blacks which is 15 points lower than that for whites does not thereby constitute a deficit.

This argument has some merit. That constellation of skills which this culture has valued highly and deemed "intelligence" is not well understood, nor are we safe in assuming that the same or essentially similar skills will continue to serve us adequately in the future. There is considerable value in diversity, especially genetic diversity. But whether or not these cognitive skills are in any sense "better" than others, or more important in the long run than certain "non-cognitive" skills which may be differently distributed, the fact remains that they appear to be the skills essential to success in scholastic endeavors and at the more prestigious and higher paying occupational levels in this society at this time. Thus the difference is clearly a deficit in terms of immediate social concerns.

Jensen's major thesis in the HER paper and in this volume is that this "difference-deficit" is most likely genetic in origin and that educational methods which rely principally on developing and using those cognitive skills in which the deficit is greatest serves only to accentuate the differences and render itself worthless. The second half of the argument, the educational implications, has been less noted and commented upon than the genetic implications, but it does deserve more thorough consideration.

Lest the incautious reader jump to the conclusions that blacks should be taught in one way and whites another, Jensen notes explicitly several times that the basis for educational decisions must be adequate and accurate psychometric assessment of all the individual's skills; e.g., "I have always advocated dealing with persons as individuals, and I am opposed to according differential treatment to persons on the basis of their race, color, national origin, or social-class background." (p. 329). Those who view all psychometric tests as nothing more than repressive tools of the establishment no doubt see such a statement as a self-serving escape clause, but the burden seems to be heavily upon such critics to demonstrate that skills so assessed are *not* educationally and socially relevant.

Jensen's principal solution is to make use of what he terms Level I abilities ("associational" cognitive skills) when Level II abilities ("analytical" cognitive skills or "intelligence") are weak or deficient. Level II abilities are required for success under traditional educa-

tional regimens, but are not necessary to learn the traditional educational content, at least through the eighth or ninth grades. Jensen further argues that such a level of achievement, considering the plight of many lower (Level II) ability children, is worth striving for.

It is obvious that many problems remain before such analyses or recommendations are acceptable, but it is to Jensen's credit that he has brought these sensitive issues into the open. Further research, especially of the educational aspects, would be most welcome. This book, as a collection of several highly relevant papers, is worth the price for those who desire more than superficial knowledge of this controversial topic.

DANIEL P. KEATING  
University of Minnesota

Melvin R. Novick and Paul H. Jackson. *Statistical Methods for Educational and Psychological Research*. New York: McGraw-Hill, 1974. Pp. XVIII + 456. \$16.50.

This is the first textbook which gives a fairly comprehensive exposition of Bayesian statistical methods directly applied to educational and psychological research. As such, it is an important book which I highly recommend.

Its importance will be felt primarily by two groups. Those who want to know what Bayesian statistics is all about and who desire concrete examples will be well served. Secondly, those of us who teach Bayesian methods, either as an integral part of a statistics course or as a special course, will find the book more than welcome. Previously, we had to rely on various references and books which had little direct application to the behavioral sciences. As a consequence, we were strong on theory and weak on real-life applications. In addition the lack of a consistent presentation relative to symbols, etc. was tough for both the teacher and student. These problems are now alleviated (if not solved). A warning to both groups is in order, however. Don't expect that you will be able to browse through the book and then appreciate the Bayesian method. Partly, this is a function of the complexity of the subject, but also because the book is not organized in such a way as to facilitate skimming. Various topics are mentioned at several places throughout the book, and one really needs to read the intervening text to fully grasp the message and/or implications. Chapter 6 entitled, "The Logical Basis of Bayesian Inference," comes closest to a "stand-alone" section which readers may skim for a quick introduction to Bayesian philosophy, but I would bet that a re-reading of this chapter, after studying the balance of the text, would give numerous additional insights.



For those who will use this as a text in their classes, he warned that the reading level is quite high, mainly due to the high information density per paragraph. Many examples are given, and conscious attempts have been made to aid the student, but even the examples must be studied very carefully in order to appreciate their import. The mathematical sophistication required is also high, but this "goes with the Bayesian territory." All in all, the teacher will have to keep his assumptions about what the students are comprehending during a reading assignment to a minimum and will have to be prepared to explain and expand upon the text. I have, in fact, spent total class periods working through examples given in the text. (This has advantages for those not happy with just lecturing.)

Some features of the book are particularly noteworthy. The emphasis on concrete applications pervades the entire text. The theme is that the methods are being presented to solve real problems, and this goes far beyond merely defining  $x$  as a test score rather than as pounds of fertilizer. Several detailed case studies, referring to published literature, make this book fairly unique in the educational statistics literature.

Several places throughout the text include sections marked off in large boxes which summarize various features of standard (and non-standard) distributions or various formulae needed to work particular prior-posterior problems. These are very helpful and will be much appreciated by students.

The set of statistical tables is excellent and is the result of a considerable amount of effort by Professor Novick and his associates. As mentioned in the text, they are a subset of a larger collection of tables. Especially noteworthy are high density intervals of the beta, inverse chi, and chi-square distributions, percentage points of the Behren's distribution and probabilities that one beta variable is larger than another. No tables of the binomial distribution are given even though the binomial is a sampling distribution featured in the text. Space is a limitation, but I would have preferred a few values of the binomial over the natural logarithm.

In this connection the text includes examples of finding binomial probabilities using the National Bureau of Standards tables, and in two of the problems at the end of the chapter, the student is asked to find probabilities using "the binomial table."

On page 116 is an example of finding a cumulative beta probability. Again, the reader is referred to tables outside of the book, i.e., Pearson's tables of the Incomplete Beta Function. The text states, correctly, that the entries are tabled for  $p \leq q$ . The example continues using interpolation in Pearson's tables and the final result is noted. It is not mentioned at all that the beta tables in the back of the book can be used for the same problem. However, these tables

are for  $p \geq q$ , which makes one wish the other tables had not been mentioned. The table entry in the book is to four decimal places without interpolation! Curious.

The writing is precise, sometimes witty and always correct. One might disagree with emphasis or interpretation, but there are no substantive errors. (At least I couldn't find any.) The book is also amazingly free of typographical errors for which the authors and publisher should be commended. (I did find one on page 244.)

The book is organized in three major sections: "Problems, Data, and Probability models," "Elementary Bayesian methods" and "Bayesian methods for comparing Parameters." The first section is an introduction, with data, that really says something, which is different from most statistics books. The second section includes analysis of binary data and the one-sample normal model. Some general theory and Bayesian concepts are included here. The third section includes the two-sample normal model, regression and correlation, and comparisons of binomial parameters.

Exercises are given at the ends of chapters. They are somewhat skimpy, especially for chapters 5 and 6, but, in general, are adequate. No answers are given.

A review is not the place to argue for any philosophy and the authors of this text do not spend much space arguing, either. They are masters of the understatement in this regard. The methods are there. Use them if you like. Some "traditionally" trained readers will be uncomfortable reading this book. This will no doubt occur when they first see the deliberate insertion of subjective beliefs into the analyses of the case studies. Their uneasiness will match the feelings I had two days before writing this review. A researcher asked me for a second class of transformations he might use so that he could obtain significance—his first tries were unsuccessful. He proudly informed me later he had found a nonparametric test that did the job—i.e.  $p < .05$ ! I respectfully ask all readers to interpret his analysis.

There is a package of computer programs which are mentioned and illustrated in the text. They are not necessary, but are very helpful. I assume one could write to Professor Novick for a listing.

My only fundamental unhappiness with the book is that  $S^2$  and  $X$ . are used for the sums of squares and the mean. Every time I make a mistake on the blackboard because of this, I will silently (or perhaps, loudly) protest. The moral is that we must all stay flexible even if it hurts—which advice I especially give to non-Bayesians who will read this book.<sup>1</sup>

DONALD L. MEYER  
University of Pittsburgh

<sup>1</sup> The Iowa Testing Program announces the publication of *Tables for Bayesian Statisticians*. (\$15.00 prepaid). A review will appear in the Summer issue of this journal.

David A. Payne, (Ed.) *Curriculum Evaluation: Commentaries on Purpose, Process, Product*. Lexington, Mass.: D. C. Heath, 1974. Pp. v & 357. \$7.95 (Paperback).

At first glance, this volume seems to be merely another of the books of readings that are in vogue these days. Closer attention shows that it is better done than many others. One of the articles was written for this book; one is a paper that seems not to have been published previously; one is an abridged composite of two that appeared in different journals; many others were abridged to suit the editor's purpose. The 44 articles are organized into five groups dealing with purposes and problems in evaluating curricula, identifying goals and objectives, design and analysis of evaluation studies, measurement techniques and problems, and illustrative evaluation projects.

In a fourteen-page "prologue," the editor has drawn on selected writings and his own experience to describe recent changes in the nature of evaluation, characterize evaluation, distinguish between evaluation and research, and characterize curriculum evaluation in particular. He has written an introduction of some two pages for each of the five parts and a paragraph introducing each of the articles. These introductions will help the reader to decide which articles may be useful.

One who is looking for objective means to measure the outcome of curriculum change will not find them in this volume. But neither will he find them elsewhere; for an area so value-laden as curriculum such means have not been devised. One looking for a model for curriculum evaluation can get some help from the articles by Barrow, Stake, and Metfessel and Michael. Perhaps the most useful parts of the volume are negative; many of the authors warn against specific pitfalls that are likely to trap curriculum evaluation. Reading the descriptions of the eleven curriculum projects provided in Part V will be a broadening experience to those who are not familiar with the quantity and variety of curriculum projects that have been developed in recent years. The essays by Tyler and Cronbach are particularly challenging in concept.

The book is a compilation of writings by persons whose experiences and ideas should prove valuable to curriculum maker and evaluators.

WILLIAM H. CARTWRIGHT  
Duke University

Joseph R. Royce (Ed.). *Multivariate Analysis and Psychological Theory*. London and New York: Academic Press, 1973. Pp. xvi + 567. £8.20 (\$23.50 in United States)

Representing a collection of the 14 papers and of the comments and rejoinders to the comments of these papers presented at the

Third Banff Conference on Theoretical Psychology, which was held from September 27 to October 2, 1971 under the cosponsorship of the Center for Advanced Study in Theoretical Psychology, the University of Alberta, Edmonton, Canada, and the Society of Multivariate Experimental Psychology (SMEP) as well as including an introductory overview paper of 13 pages by the editor, this volume reflects the basic view that theory construction in psychology can be greatly facilitated through the application of multivariate research strategies, particularly factor analysis. The contributions of the distinguished participants have been organized into two major divisions: (a) "Part I, Methodological, Pre-Theoretical and Meta-Theoretical Issues" and (b) "Part II, Toward a Comprehensive, Multivariate Psychological Theory." To illustrate the issues of Part I, an enumeration of the titles of the eight papers and their authors would seem helpful: "Right Answers to the Wrong Questions? A Re-examination of Factor Analytic Personality Research and its Contribution to Personality Theory," K. Pawlik; "Linear Regression Equations as Behavior Models," K. V. Wilson; "How shall We Conceptualize the Personality We seek to Investigate?" D. W. Fiske; "Prescriptions for a Multivariate Model in Personality and Psychological Theory: Ecological Considerations," S. B. Sells; "Multivariate Approaches to the Study of Cognitive Styles," P. E. Vernon; "Comparative Studies of Multiple Factor Ability Measures," S. G. Vandenberg; "Theory of Functions Represented among Auditory and Visual Test Performances," J. L. Horn; and "Theoretical Issues and Operational-Informational Psychology," J. P. Guilford. Similarly for Part II, the six titles and corresponding authors were as follows: "Multivariate Models of Cognition and Personality: The Need for Both Process and Structure in Psychological Theory and Measurement," S. Messick; "The Conceptual Framework for a Multi-Factor Theory of Individuality," J. R. Royce; "Causal Theories of Personality and How to Test Them," J. A. Gray; "Key Issues in Motivation Theory (with Special Reference to Structured Learning and the Dynamic Calculus)," R. B. Cattell; "A Multidimensional Theory of Depression," T. Weckowicz; and "The Psychological Structure of Peer Group Forces in Delinquency," D. S. Cartwright and Katherine Howard.

In setting the stage for what is to follow Royce describes the current state of the art in multivariate theoretical psychology, differentiates between the terms "multivariate" and "theory," reviews several of the technical problems and issues underlying factor analysis relative to its use in the development of a comprehensive, multivariate psychological theory, discusses briefly issues in scientific methodology as applied to psychology, and endeavors to explain the role of theory in psychology in relation to each of four



stages through which advanced scientific disciplines have gone. Next, he presents a succinct overview of the conference papers from which the reader can readily grasp both the basic structure of the volume and the principal methodological and substantive issues considered. From this excellent integrative summary the reader can probably decide which papers he may wish to study in detail.

In Part I the first two papers by K. Pawlik and K. V. Wilson tend to be concerned very much with the underlying tenets and applicability of the factor analytic model and related linear regression models to theory construction. The authors as well as the discussants give considerable attention to nonlinear and interactive as well as linear components manifest in behavioral data and to the need for appropriate modifications in the familiar compensatory linear model to permit an improved conceptualization and understanding of behavior patterns. Somewhat more substantively oriented than are the first two papers, the third contribution by Fiske deals with how to conceptualize and measure personality first through identifying and examining several critical problems and then by proposing steps to resolve these problems, although several problems are left as issues to be considered. Emphasis is placed upon the need to define constructs in both their behavioral and contextual (situational) facets as well as upon extensive psychometric efforts in a variety of settings. In the fourth paper Sells endeavors to develop a comprehensive multivariate model of personality that embraces personal and environmental components with the view of identifying sources of variance attributable to their subcomponents. As its main thrust the personality model would be toward constructing a functional system of components that would account for the major operations of the organism and that would eventually provide a means for viewing the patterns of the operational capacities of these components as a basis for a typology of personality. Subsequent to the fifth paper by Vernon, which is a scholarly but relatively short historic review of cognitive styles, the sixth and relatively long paper by Vandenberg is a massive effort to present critically relevant data from numerous studies concerning multifactor ability measures for evaluating their usefulness relative to seven criteria. In the seventh paper, which is quite specific in its concern with the development of a theory of psychological function of auditory tests, Horn has amassed numerous data from his own studies and those of several other investigators and has endeavored to interrelate auditory and visual abilities. As the eighth and concluding paper in Part I, the one by Guilford on theoretical issues associated with an "operational-informational" point of view is an examination of the relationships between his approach and each of five historical-theoretical issues: mental faculties vs. mental factors, mental acts vs. mental contents,



difficulties in elementarism (e.g., simple behavior units vs. Gestalt configurations), the shortcomings of associationism (including behaviorism), and subjective vs. objective orientations to the study of human behavior. In essence Guilford has tried to develop a general theory of behavior which clearly has its roots in the Structure-of-Intellect model.

Although the papers in Part I supposedly emphasize "meta-theoretical, methodological, and pretheoretical [primarily empirical or minimally substantive theory oriented]" concerns, the six papers in Part II have been categorized as mainly substantive-theoretical in that they tend to go beyond the data and to afford a speculative explanatory basis for understanding the complexities of human behaviors. As the first of six papers in Part II, the one by Messick is an attempt to take into account both structural and functional (dynamic) components of behavior. After reviewing Guilford's structure-of-intellect operational-information model and hierarchical models of intellect, Messick proposes the need for a model that considers the sequence of operations in complex cognitive processes as in learning and concept attainment, perception and attention, memory and recall, and problem solving and creativity, and then goes on to consider stylistic aspects of cognition as reflecting personality dimensions "that cut across affective, personal-social, and cognitive domains and thereby serve to interweave the cognitive system with other subsystems of personality organization [p. 287]." In his comprehensive formulation Messick also considers developmental changes in cognition as they interact with environmental variables. Similarly in the second paper Royce expounds a general theory of individual differences relative to which he sets forth a hierarchical model for affective, cognitive, and style structures. He endeavors to incorporate process or change characteristics within his theory through examining the ontogeny of factors, pointing out hereditary and environmental sources of variation, posing a factor-gene model, considering cultural-learning mechanisms in relation to factors, and finally speculating upon the neural mechanisms underlying cognitive and affective processes in individuality.

Continuing with a physiological-neurological emphasis, Gray, in the third paper, presents a causal theory of personality in which he attempts to specify three different brain systems as essentially isomorphic with corresponding manifestations of temperament-emotionality as conceptualized with some modifications in Eysenck's familiar dimensions (introversion-extroversion, neuroticism, and psychoticism). Evidence relevant to the theory is presented from experiments of children's behaviors in an operant conditioning task. Also interested in emotionality and the affective side of human behaviors, Cattell in the fourth paper of Part II endeavors to define

motivation measurement by offering a familiar factor specification equation embodying ability, temperament, and dynamic components, of which the third mentioned one includes a modulating index that reveals by how much an ambient (existing general background) stimulus provokes an increase in a dynamic trait. In his discussion of his theory Cattell directs his attention to a number of concerns about which he expounds at length: (1) the riddle of the nature of the components of motivation, (2) the measurement of conflict in the clinical field and the prediction of decision, (3) the identification of extra-motivational determiners of conflict involving possible relationships between measurable personality factors and generalized dynamic factors, (4) the introduction of three classes of measurable predictors within the context of learning theory (general psychologic states, ergic tension levels, and reward as tension reduction) with particular emphasis on emotional learning, and (5) the need for experimentation to test the workability of the principles or parameters set forth in the structured learning theory.

The two concluding papers in Part II tend to be rather specific in emphasis in that the fifth one by Weckowicz is concerned with how multivariate concepts and strategies could be useful in delineating several of the theoretical and practical issues in psychopathology as in the development of a multidimensional theory of depression and in that the sixth and concluding contribution by Cartwright and Howard is directed toward the development of a multivariate model to describe peer-group forces in delinquency. These last two papers serve to illustrate both the power and versatility of multivariate methodology as a tool in aiding the psychologist to work in two distinct areas of great social importance.

The evaluation of a collection of papers arising from a conference is a difficult task especially when the participants have been given substantial freedom in the selection and preparation of their papers, for the three broad criteria set forth at the time of the call for papers were that each contribution would be multivariate, substantive, and theoretical in scope. Despite the differing emphases in the papers, the editor succeeded in organizing and sequencing them in such a way as to achieve a relatively high degree of coherence if not unity in the final product. An additional unifying agent has been the insertion of an introductory section as well as the conclusion of comments from one or more of the conference participants and of rejoinder statements from each author. Hence this substantial interaction among the membership of the conference has had an integrating as well as a clarifying influence that has tended to reconcile similarities and differences in points of view expounded.

There is little doubt that this comprehensive volume represents

much of the latest thinking of several of the most distinguished contemporary theorists in psychology. It affords numerous suggestions—either direct or implied—for needed research efforts at a methodological, theoretical, and empirical level. In addition to its broad coverage, it makes evident the rich potential of multivariate approaches to theory building and to the advancement of psychology as a science. Important fringe benefits include the extensive bibliographies in several of the chapters, an author index, a detailed subject matter index, and a workable table of contents.

In summary this volume is a significant contribution to the theoretical literature in psychology. It is also of great importance to the psychometrician who can gain improved insight regarding the contributions which multivariate analysis can make to the conceptualization and understanding of the complexities of human behavior. Both the psychological theorist and methodologist should have this highly sophisticated and thought-provoking book as an integral part of their professional libraries.

WILLIAM B. MICHAEL

*University of Southern California*

Julian C. Stanley, Daniel P. Keating, and Lynn H. Fox (Eds.).  
*Mathematical Talent: Discovery, Description, and Development.* Baltimore, Md.: Johns Hopkins University Press, 1974.  
Pp. xvii + 215. \$10.00 cloth, \$2.95 paper.

The nine research and discussion papers contained in this volume are the consequence of a five-year project, the Study of Mathematically and Scientifically Precocious Youth (SMSPY), sponsored at The Johns Hopkins University by the Spencer Foundation. Julian Stanley, the director of the project, sets the stage in Chapter 1, "Intellectual Precocity," with a brief review of earlier work on talented youth. The book is dedicated to Lewis Terman, and his classic study of the gifted, in spite of its recognized flaws, is highlighted. Several case histories of mathematically gifted young people are also discussed.

Stanley writes so well that the reviewer found himself desiring a more comprehensive discussion instead of the short shrift to which he was treated in Chapter 1. Although Stanley is justifiably critical of society's failure to support research and education of the gifted while providing funds aplenty for the educationally disadvantaged, it should be noted that the disadvantaged are greater in number and a more serious economic problem than the gifted. Whether the gifted would sufficiently fulfill their potentialities without special treatment is a question unanswered.

Chapters 2 through 4 describe the major findings of the project after its first year of operation. The data obtained on 35 mathe-



matically precocious boys and eight mathematically precocious girls identified in a mathematics and science competition participated in by 396 seventh and eighth graders are described by Donald Keating in Chapter 2. A number of interesting findings emerged from the tests and questionnaires administered: pronounced sex differences in favor of the boys, disillusionment with school on the part of many gifted students, and no indication of personality difficulties in the gifted boys. Other findings, e.g. the higher educational level of the parents and greater investigative interests in their gifted children are less surprising. Furthermore, the range of inter- and intraindividual variability on the test and questionnaire data are considerable. By way of criticism, it is unfortunate that no interviews or observational data of the sort provided by Krutetskii (1966) on younger children were collected.

Continuing with the four core chapters, in Chapter 3 Lynn Fox discusses alternative ways of facilitating the development of mathematically precocious youth—special schools, enrichment, acceleration, and early admission to college or specific college courses. Based on the SMSPY data, Fox makes a strong case for individualizing instruction of the gifted, particularly by letting them take college courses early. In Chapter 4 Helen Astin discusses the reportedly unanticipated finding of significant sex differences in favor of boys obtained in the project. Not only did the boys perform at a higher level than the girls, but the former reported less liking for school, showed an interest and precocity in mathematics at an earlier age, tended to be the oldest children from large families, but showed less tenderness, sympathy, and sociability than girls. Both gifted boys and girls tended to come from achievement-oriented, middle-class families and to be rated as likeable children by their parents but to show less tenderness, sympathy, and sociability than girls.

Anne Anastasi's commentary on the precocity project, given in Chapter 4, brings out some of its strengths and weaknesses and makes a number of valuable suggestions for further research. In addition to being careful with the use of terms such as "latent talent," studies of the childrearing practices of parents of the gifted, and the role of social encouragement and special materials in fostering sex differences in abilities are some of the recommendations made by Anastasi.

Chapters 6 through 9 were written especially for this volume, and form a compendium of "afterthoughts" concerning the project. In Chapter 6, Fox describes the results of a special accelerated mathematics program participated in by 19 gifted students who had just completed sixth grade. By means of guided independent study, in a very short time these students learned algebra. Analysis of test scores and other data obtained from these students revealed the importance of the parents' interests and motivation, and other

(verbal) aptitudes as well as mathematical ability in promoting success in this subject. Social factors are especially important for girls, and interest is more important in determining perseverance than in actual achievement.

Along with ability tests and other measures, the California Psychological Inventory (CPI) was administered to the 35 mathematically gifted boys in the main study. The CPI profiles of this group were compared with those of an eighth grade random, an eighth grade gifted, a high school gifted, and a high school norm group. Relying on these data, the authors of Chapter 7 criticize the "traditional assumptions" that the gifted are interpersonally ineffective or maladjusted. Although the authors employed some questionable statistical procedures and tend to jump to conclusions somewhat, they make their points that the gifted are not generally maladjusted and that acceleration would not have deleterious effects on their personality development.

The career interests and values of the gifted are discussed in Chapter 8, based on administration of the Vocational Preference Inventory (VPI) and the Study of Values. The findings that junior high students who enter a mathematics and science competition have interests in mathematical and scientific occupations are not surprising, but some of the specifics, such as the fact that boys more than girls prefer investigative occupations, are worth noting. Sex differences were also found on the Study of Values scales. The theoretical value was highest for most of the 35 male competition winners, and the girls' social values were higher than those of the boys. The point is again made that academic ability alone, in the absence of appropriate interests and evaluative attitudes, does not always lead to precocious achievement. It is unfortunate that the brief measures of interests and values used here did not tap these variables better. The Strong Vocational Interest Blank, a depth interview, a good biodata inventory, and careful observations of gifted students would have provided more information on their interests and personalities.

The last chapters of the volume give the results of an assessment of the study habits and attitudes, observations of classroom behavior, and teacher and student impressions of five junior high school boys who took a college course in mathematics. The fact that these students did well in the class and were assimilated successfully lends credence to the notion that acceleration of gifted children is the best policy, especially when there are several such children together.

In an epilogue summary of Chapters 6-9, the editors of the volume recapitulate the purposes of the project and some of the questions that were examined. Although many of these questions concerning the development of mathematical ability, its effects on



social and emotional development, and the origins of individual and sex differences require more extensive study, the results reported here permit at least tentative conclusions on several matters. Precocious students' social and emotional development does not appear to be unduly harmed by separating them from their age group, and they benefit intellectually by being accelerated.

Although the Study of Mathematically and Scientifically Precocious Youth is not of the same magnitude as Lewis Terman's landmark longitudinal investigation, it represents a much-needed reopening of issues concerning the development and training of individuals possessing special abilities. In addition to mathematical precocity, other abilities demand continued study. As noted in this volume, SMSPY is being complemented by a parallel investigation of verbally precocious children.

There are so many unanswered questions on the psychology and education of human abilities that it is difficult to say what we know for certain and what problems would be worth investigating. But in spite of small sample sizes, inadequate instruments, and questionable methodology in certain instances, the Stanley, Keating, and Fox study provides an impetus and some direction. Perhaps its most meritorious contribution, like that of Lewis Terman, is to be found in its failure to support the persisting myths about the social and emotional development of the mathematically gifted and the perils of academic acceleration. Otherwise, as with almost any ambitious research project in psychology, it poses more questions than it answers.

LEWIS R. AIKEN

John A. R. Wilson, Mildred C. Robeck, and William B. Michael.  
*Psychological Foundations of Learning and Teaching*. (2nd ed.)  
New York: McGraw-Hill, 1974. Pp. xv & 589. \$10.95 and \$8.95  
(paperback).

*Psychological Foundations of Learning and Teaching* is an outstanding text in the field of educational psychology. These chapter titles are indicative of its scope "1. Motivation for better teaching, 3. Appraisal and objective of education, 4. Forming cognitive associations, 5. Affective associations, 7. Affective conceptualizations, 9. Neurology of learning, 12. Development of perceptual abilities, 14. Cognitive growth: Piaget's theory, 15. Intelligence: structure and function, 17. Evaluation of objectives, 18. Teacher-made tests and scales, 19. Statistical methods." The book concludes with a 22 page bibliography. Largely through the courtesy of the United Nations, students of different lands are shown in learning situations. Each chapter begins with a list of behavioral objectives which imply the abilities to be acquired, thought questions to be answered,

and technical terms to be defined and used. For example, Chapter 6 Cognitive Conceptualizations: Grasping Inherent Relations begins with the following list of objectives or goals:

This chapter is planned to help you structure curriculum so that your students will learn relationships that are transferable to new situations. You will be able to:

Observe when a student synthesizes knowledge, or associations, to form a conceptualization.

Plan a series of questions that point to conceptualizations.

Design a series of steps that lead to discovery.

Design learning situations to provide checks against which conceptualizations can be tested.

Analyze a unit of content into essential features.

Conceptualize the following terms:

intuition

inquiry process

gestalt

insight

intrinsic programming

inductive learning

Study of this text should acquaint the student with the contributions of Benjamin Bloom, Sigmund Freud, J. P. Guilford, B. F. Skinner, L. L. and Thelma Thurstone, Robert Thorndike, Ralph Tyler, and many others. Both the Cognitive and Affective Domain Taxonomies are carefully described and illustrated. Similarly, Guilford's model for the structure of intellect and the correlations of the Thurstone primary mental abilities with each other and with total scores are also explained and illustrated.

Chapter 17 Evaluation of Objectives includes discussion of educational accountability, reliability, and the different kinds of validity—content, criterion related, and construct. Chapter 18 explains the construction and use of teacher-made and standardized tests and scales. A number of series of exercises are quoted from a folio of Chicago City Junior College, English and General Course Examinations. Chapter 19 Statistical Methods covers most of the important procedures of descriptive and inferential statistics.

This admirable book is well worth using as a text and as permanent addition to ones professional library.

MAX D. ENGELHART

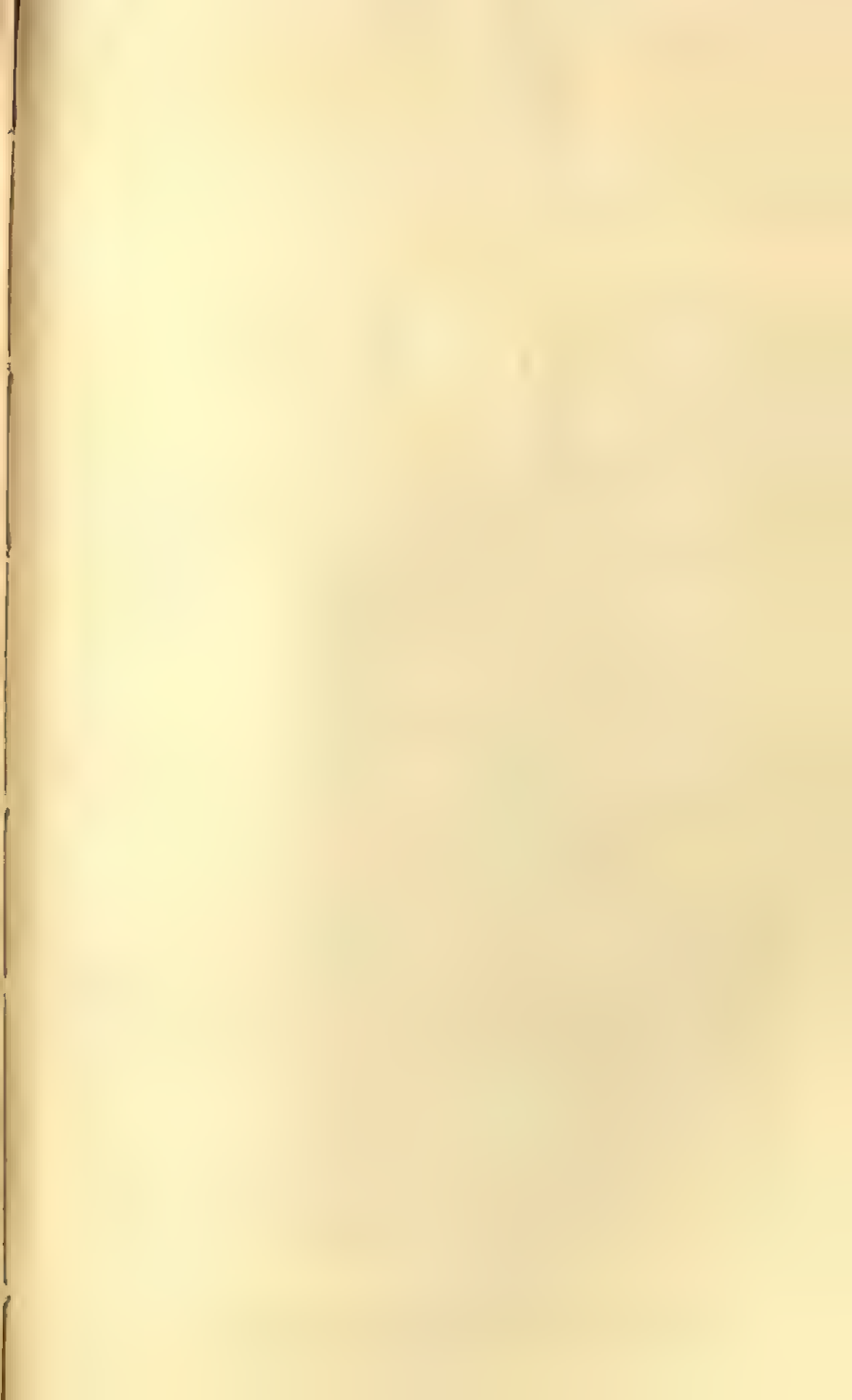
## ERRATUM

In E. B. Page's article "Top-down" Trees of Educational Value, which appeared in the Autumn of 1974 issue pp. 573-584, on p. 579, formula (2) should read:

$$R = \sqrt{\sum_{i=1}^n \beta_i^2} = \sqrt{\sum_{i=1}^n r_{ic}^2} . \quad (2)$$

Thus the variance explained,  $R^2$ , would be calculated by summing the *squared* correlations (Page and Breen, 1973). The writer regrets this typographical omission, which was brought to his attention by Professor Jason Millman. The logic of the article, and of the subsequent formulas, is unaffected by the change.









EDUCATIONAL and  
PSYCHOLOGICAL



# MEASUREMENT

Diary No. 935 .....

Date 28/8/75 .....

il. N SKB. ....

Bureau Educ. Psy. Research.

W. SCOTT GEHMAN, *Editor*

GERALDINE R. THOMAS, *Managing Editor*

WILLIAM B. MICHAEL, *Editor, Validity Studies and Computer Programs*

JOAN J. MICHAEL, *Assistant Editor, Validity Studies and Computer Programs*

MAX D. ENGELHART, *Book Review Editor*

LEWIS R. AIKEN, JR., *Assistant Book Review Editor*

FREDERIC KUDER, *Editor Emeritus*

## BOARD OF COOPERATING EDITORS

DOROTHY C. ADKINS, *University of Hawaii*  
LEWIS R. AIKEN, JR., *Guilford College*  
HAROLD P. BECHTOLDT, *The University of Iowa*  
WILLIAM V. CLEMANS, *American Institutes for Research*

LOUIS D. COHEN, *University of Florida*  
ANTHONY J. CONGER, *Duke University*  
JENNIS A. DAVIS, *Research Triangle Institute*  
HAROLD A. EDGERTON, *Performance Research, Inc.*

GENE V GLASS, *University of Colorado*  
J. P. GUILFORD, *University of Southern California, Los Angeles*

JOHN A. HORNADAY, *Babson College*  
JOHN E. HORROCKS, *The Ohio State University*  
CYRIL J. HOYT, *University of Minnesota*  
MILTON D. JACOBSON, *University of Virginia*  
JOSEPH C. JOHNSON II, *Jackson State University*

WILLIAM G. KATZENMEYER, *Duke University*  
ROBERT E. LANA, *Temple University*  
FREDERIC M. LORD, *Educational Testing Service*  
ADDIE LUBIN, *Navy Medical Neuropsychiatric Research Unit, San Diego*

LOUIS L. MCQUITT, *University of Miami, Coral Gables*

HOWARD G. MILLER, *North Carolina State University at Raleigh*

ROBERT L. MORGAN, *North Carolina State University at Raleigh*

HENRY MOUGHAMIAN, *City Colleges of Chicago*

DAVID NOVAK, *The Neuse Clinic, New Bern, N. C.*

ELLIS B. PAGE, *University of Connecticut*

NAMBURY S. RAJU, *Science Research Associates, Inc.*

BEN H. ROMINE, JR., *University of North Carolina at Charlotte*

THELMA G. THURSTONE, *University of North Carolina at Chapel Hill*

WILLARD G. WARRINGTON, *Michigan State University*

JOHN L. WASIK, *North Carolina State University at Raleigh*

KINNARD WHITE, *University of North Carolina at Chapel Hill*

JOHN E. WILLIAMS, *Wake Forest University*

E. G. WILLIAMSON, *University of Minnesota*

VOLUME THIRTY-FIVE, NUMBER TWO, SUMMER 1975



## CONFIGURAL FREQUENCY ANALYSIS AS A STATISTICAL TOOL FOR DEFINING TYPES

G. A. LIENERT AND J. KRAUTH

University of Dusseldorf

Configural frequency analysis (CFA) is a new method for identifying types. Types are defined as patterns (configurations) of binary variables occurring more frequently than may be expected under the assumption of complete independence of the respective variables, and are tested for significance by multiple binomial tests or suitable approximations. CFA is illustrated numerically by an example. Relations to latent class analysis and to factor analysis are discussed. It is suggested to use CFA as a type-defining method instead of factor analysis if the variables are linked to each other not only by first but also by higher-order associations.

There are many definitions of the concept of type (see Cattell, Coulter, and Tsujioka, 1968), no one being satisfactory from a statistical point of view. In fact, up to now a type has been considered "a statistical concept without statistics" (English and English, 1957).

Since intuitively a type is conceived as a pattern of qualities that tend to occur together with high frequency (see Lorr, 1966) a type might tentatively be defined as a modal frequency in a discrete multivariate distribution.

However, modal frequencies may occur from two different sources, i.e., (a) from frequently occurring component qualities, and (b) from interactions which increase frequencies of the component qualities.

While source (a) is trivial in producing modal frequencies, source (b) is suggested to define a type modal frequency as follows: A type is a multivariate class of qualities that occur more often together than may be expected by chance from the proportions of the respective qualities under the assumption of their independence.

For example, extroverted smoking men define a type if and only if

they occur more often in a population than may be expected by chance from the proportion of extroverts, from the proportion of smokers, and from the proportion of men, in that population.

According to the definition above, a method will be proposed for identifying types statistically. Originally introduced as a heuristic method (Lienert, 1968) the so-called configural frequency analysis (C-FA) recently has been developed into an inferential method (Krauth and Lienert, 1972).

### *Rationale of CFA*

If  $t$  binary variables ( $v = +, -$ ) are observed in a sample of  $N$  individuals,  $2^t$  binary classes or configurations  $C_j, j = 1, 2, \dots, 2^t$ , occur. The  $C$ 's define a  $t$ -dimensional fourfold table with *observed* configural frequencies,  $f$ , or configural proportions,  $\hat{p}$ , where

$$\hat{p} = f/N. \quad (1)$$

Each of the  $2^t$  observed frequencies is associated with an *expected* frequency  $e$ , or proportion  $P$ , where

$$P = e/N. \quad (2)$$

Under the null hypothesis of no interaction of any of the  $t$  variables, i.e., under the no-type hypothesis there is

$$p = P \quad \text{for all } C_j\text{'s}, \quad (3)$$

while under the alternative (one-sided) hypothesis of interaction, or type hypothesis there is

$$p > P \quad \text{for at least one } C_j. \quad (4)$$

For testing whether the null hypothesis ( $H_0$ ) holds, the model must be specified for getting expected proportions.

1. According to the *fixed* proportions model where the proportions  $\pi_{iv}$  of the  $t$  variables or their binary classes are unknown or postulated theoretically,  $P = \pi$  is given by

$$\pi = \prod_{i=1}^t \pi_{iv}, v = +, -. \quad (5)$$

2. In the estimated proportions model the proportions  $P$  are estimated from the *variable* proportions  $\hat{p}_{iv}$  of the sample by

$$P = \prod_{i=1}^t \hat{p}_{iv}, v = +, -. \quad (6)$$

In general, model 2 is more realistic than model 1, since fixed



proportions are seldom known except in the case where standardized test scales have been dichotomized at their population medians to give  $\pi_{lv} = 1/2$ . Therefore, in the following, only model 2 will be considered for testing.

### *Binomial Testing for Types*

When associated with an expected proportion  $P$  the probability under  $H_0$  that a specified configuration  $C$  has an observed frequency greater or equal  $f$  is given by

$$\text{Prob}(C) = \sum_{i=f}^N \binom{N}{i} P^i (1-P)^{N-i}. \quad (7)$$

A configuration  $C$  is assumed to be a type if  $\text{Prob}(C) \leq \alpha$ , where  $\alpha$  is the significance level agreed upon prior to sampling, and valid for only one test.

If, as usual in heuristic exploration,  $2 \leq r \leq 2^t$  configurations are tested for types,  $\alpha$  must be adjusted for multiple testing. As Krauth and Lienert (1972) have shown, adjustment is most simply made by setting

$$\alpha^+ = \alpha/r. \quad (8)$$

Of course, specified testing of selected  $r$  out of  $2^t$  configurations has to be justified prior to sampling while unspecified testing for all  $2^t$  configurations is not restricted to justification.

### *Approximate Testing for Types*

1. If  $N$  is large and  $P$  is not too small, *normal* approximation to the binomial may be made by setting

$$z(C) = (f - NP - .5) / \sqrt{NP(1-P)}. \quad (9)$$

$H_0$  is rejected, if  $z(C) \geq z^+$ , where  $z^+$  is the unit normal deviate associated with  $\alpha^+$ .

2. Instead of the normal approximation, provided  $NP > 5$  for all  $r$  configurations to be tested, the *chi-square* approximation

$$X^2(C) = (f - NP)^2 / NP, \quad df = 1 \quad (10)$$

may be used. However, since chi-square is sensitive to the alternative  $p > P$  as well as to the alternative  $p < P$ ,  $X^2(C)$  tests for types as well as for "antitypes" if  $X^2(\alpha^+)$  is set as the critical limit. For testing against  $p > P$  only,  $X^2(2\alpha^+)$  is the correct limit. For unconventional  $\alpha^+$ 's, use the relation  $X^2 = z^2$  valid for  $df = 1$ .

3. If  $N$  is very large and  $P$  is very small, *Poisson*-approximation to the binomial is most suitable. Further approximations are reviewed by Molenaar (1970).

4. In every case the binomial fractiles may be calculated through the *F*-distribution by using the *Camp-Paulson* approximation (see Molenaar, 1970)

$$F(C) = \frac{(f+1)(1-P)}{(N-f)P} \quad (11)$$

$F(C)$  is to be evaluated for  $df_1 = 2(N-f)$  and  $df_2 = 2(f+1)$  with  $F(\alpha^+)$  as a critical limit. For unconventional  $\alpha^+$ 's, the critical limit may be obtained by Paulson's normal transformation of the *F*-distribution (see Kendall and Stuart, 1969).

### *An Example from Experimental Psychopathology*

$N = 65$  volunteer subjects ( $S$ s) were rated for occurring (+) or not occurring (-) of the symptoms  $H$  = Hallucinations,  $B$  = Black-outs,  $T$  = Thinking disturbances, and  $A$  = Affective reactions, under lysergic acid diethylamide (LSD). The configural frequencies  $f$  of the  $t = 4$  binary symptoms are given in Table 1 (Data from Lienert 1970).

1. The estimations of the expected proportions,  $P$ , in Table 1 were calculated (a) by counting the four pairs of one-dimensional marginals

TABLE 1

HBTA	$f$	$P$	Prob(C)	$z$	Prob( $z$ )
++++	12	.07696	.00371	3.024	.00124
+++-	0	.04214	1.00000	-2.000	.97725
++-+	1	.07017	.99117	-1.972	.97570
++--	4	.03842	.24969	.647	.25872
+--+	1	.05824	.97976	-1.740	.95907
+--+	3	.03189	.34333	.302	.38133
+---	5	.05310	.26253	.580	.28096
----	0	.02908	1.00000	-1.764	.96113
-+++	8	.11544	.48038	.001	.50040
-++-	1	.06322	.98566	-1.840	.96712
-+-+	3	.10525	.97285	-1.755	.96037
-+--	8	.05764	.03286	1.998	.02286
--++	2	.08736	.98103	-1.835	.96675
--+-	7	.04784	.03542	1.970	.02442
---+	10	.07965	.03231	1.980	.02385
----	0	.04362	1.00000	-2.025	.97857

CFA of  $t = 4$  binary symptoms  $H$  = Hallucinations,  $B$  = Black-outs,  $T$  = Thinking disturbances and  $A$  = Affective reactions of  $N = 65$  volunteer  $S$ s under influence of lysergic acid diethylamide.



selecting  $s$  out of  $t$  variables CFA may be performed hierarchically. Referring to the example on LSD, the instructions are:

1. Perform a CFA in Table 1 with four symptoms and evaluate  $\sum z^2$  like a chi-square for  $2^4 - 4 - 1 = 11$  df, the  $z$ 's being the unit normal deviates associated with the *Prob*'s.
2. Perform CFA's of all  $\binom{4}{3} = 4$  triplet combinations of symptoms (*HBT*, *HBA*, *HTA*, *BTA*) and evaluate each  $\sum z^2$  for  $2^3 - 3 - 1 = 4$  df.
3. Finally perform CFA's of all  $\binom{4}{2} = 6$  doublet combinations of symptoms (*HB*, *HT*, *HA*, *BT*, *BA*, *TA*) and evaluate each  $\sum z^2$  for  $2^2 - 2 - 1 = 1$  df.
4. Now select that CFA out of all 11 CFA's which gives the "most significant"  $\sum z^2$ .

Proceeding that way with the symptoms in Table 1, the triplet *BTA* gives the "most significant" chi-square, and the most clear-cut syndrome type  $B + T + A +$ . In general, hierarchical CFA is a means for eliminating variables irrelevant in defining configural types.

### *CFA and Related Type-defining Methods*

CFA is formally and substantially linked to Lazarsfeld's Latent Class Analysis (LCA) and to the well known factor analysis (FA).

1. LCA starts, as CFA, from a pattern of binary variables (see Lazarsfeld and Henry, 1968, or Cassady, Miller, and Dingman, 1968) but arrives, opposite to CFA, at types of statistically independent variables similar to modal frequencies of source (a) mentioned initially. Furthermore, LCA types are conceived as points along a one-dimensional continuous space while CFA types are points within a  $t$ -dimensional binary space. Thus, LCA and CFA are incompatible type-defining methods, though neither method is restricted to first-order associations between variables.

2. FA is primarily related to hierarchical CFA in so far as both methods reduce the number of variables to those relevant for defining types or factors. FA differs from CFA in that FA relies only on first order associations between binary variables while CFA relies on first and higher-order associations. Thus FA and CFA give comparable results only if there are no higher-order associations in binary variables or hypernonlinear correlations in (dichotomized) continuous variables. Both LCA and FA differ from CFA in that (a) CFA is a type-defining method giving unique solutions even as a heuristic procedure, (b) CFA is, unlike LCA and FA, an inferential method and (c) CFA is completely nonparametric while FA is parametric and LCA has parametric implications. The only disadvantage of CFA is that the sample size is required to increase exponentially as the number of variables increases linearly.

### Conclusions

CFA is suggested as a type-defining method primarily for variables related to each other not only by first-order associations, but also by second-order and higher-order associations. In case of the LSD example, Black-outs, Thinking disturbances and Affective reactions are linked to each other by a second-order association (see Goodman, 1964), and the type  $B + T + A +$  could never have been isolated by means of FA or any other method of intercorrelation analysis.

Higher-order associations and hypernonlinear correlations seem to occur mostly in psychopathology as has been shown for depressive symptoms (Lienert, Angst, Baumann, Gebert, 1973) and for aphasic test scales (Gloning, Lienert, Quatember, 1972). It is, therefore, suggested that clinical syndromes and types of personality disorders may be re-examined by a CFA of their symptoms or traits, especially if they have failed to be identified conclusively and consistently by other type-defining methods.

CFA is suggested as the only valid type-defining method, if the variables, or some of them, are scaled nominally and multinary. CFA of multinary variables is as straightforward as CFA of binary variables, if binomial tests or their chi-square approximations are used to compare observed with expected configural frequencies. Of course, the number of configurations possible will be enlarged in case of multinary variables thus requiring a larger sample size for efficient binomial testing.

### REFERENCES

- Cassady, J. M., Miller, C. R., and Dingman, H. F. Latent class analysis: A direct approach. *APA-Proceedings*, 1968, 76, 209-210. New York: APA-Press.
- Cattell, R. B., Coulter, M. A., and Tsujioka, B. The taxonometric recognition of types and functional emergents. In R. B. Cattell (Ed.) *Handbook of multivariate experimental psychology*. Chicago: Rand McNally, 1968, Ch. 9.
- English, H. B. and English, A. C. *A comprehensive dictionary of psychological and psychoanalytical terms*. New York: Longmans, Green and Co., 1957.
- Gloning, K., Lienert, G. A., and Quatember, R. Konfigurationsfrequenzanalyse aphasie-spezifischer Testleistungen. *Zeitschrift für klinische Psychologie und Psychotherapie*, 1972, 17, 115-122.
- Goodman, L. A. Simple methods for analyzing three-factor interaction in contingency tables. *Journal of the American Statistical Association*, 1964, 59, 319-352.
- Kendall, M. G. and Stuart, A. *The advanced theory of statistics* I. London: Griffin, 1969, 3rd Ed.
- Krauth, J. and Lienert, G. A. Nichtparametrischer Nachweis von Syndromen durch simultane Binomialtests. *Biometrische Zeitschrift*, 1973, 15, 13-20.



- Krauth, J. and Lienert, G. A. Konfigurationsfrequenzanalyse, Freiburg/Brsg., Verlag Karl Alber, 1973.
- Lazarsfeld, P. F. and Henry, N. W. *Latent structure analysis*. Boston: Houghton-Mifflin, 1968.
- Leuner, H. *Die experimentelle Psychose*. Heidelberg: Springer, 1962.
- Lienert, G. A. Die 'Konfigurationsfrequenzanalyse' als Klassifikationsmittel in der klinischen Psychologie. In Irle, M. (Hrsg.) *Bericht über den 26. Kongress der Deutschen Gesellschaft für Psychologie*, 1968. Göttingen: Hogrefe, 1969, 244-253.
- Lienert, G. A. Konfigurationsfrequenzanalyse einiger Lysergsäure-diäthylamid-Wirkungen. *Arzneimittelforschung* (Drug research), 1970, 20, 912-913.
- Lienert, G. A., Angst, J., Baumann, U., and Gebert, A. Association structure analysis of depressive symptoms. *Therapiewoche*, 1973, 23, 1-6.
- Lorr, M. *Explorations in typing psychotics*. Oxford: Pergamon Press, 1966.
- Molenaar, W. *Approximations to the poisson, binomial and hypergeometric distribution functions*. Amsterdam: Mathematical centre, 1970.

## A METHOD FOR HIERARCHICAL CLUSTERING OF A MATRIX OF A THOUSAND BY A THOUSAND<sup>1</sup>

LOUIS L. MCQUITTY AND VALERIE L. KOCH

University of Miami, Coral Gables

Most hierarchical clustering methods are limited to relatively small matrices. On the other hand, if personality types exist, there are probably many of them. Consequently, a method is needed for clustering hierarchically the interrelationships between many persons, as represented in a matrix of a thousand by a thousand. This paper develops and illustrates such a method.

At least two previous studies indicate that if personality types exist, there are many of them (McQuitty, 1954 and 1957). If large numbers of types do in fact exist, studies based on sample sizes which do not adequately represent them, could fail to yield them just because they are based on too few cases.

Investigations of whether or not personality types exist would be facilitated by a method which clusters hierarchically a matrix reporting interassociations between a thousand or more persons. This paper describes one such method. The method is both concise and rapid.

### *Method*

#### *Initial Definitions of Types*

The method of this paper is developed out of two definitions of types: (1) reciprocal, dyadic types and (2) higher order types. A reciprocal, dyadic type is defined in relation to a set of objects; it in-

<sup>1</sup> This investigation was supported in part by Public Health Service Research Grant No. MH 14070-03 from National Institute of Mental Health.

cludes only two objects,  $O_i$  and  $O_j$ ; Object  $O_i$  is most like Object  $O_j$ , and Object  $O_j$  is in turn most like Object  $O_i$ , i.e., amongst the objects of the set.

Dyadic types (themselves reciprocal pairs) are associated by reciprocal pairs into higher order types. In brief, this is accomplished by dropping temporarily one member of each dyadic type and thus allowing the remaining members to form new reciprocal pairs. When the members of the new reciprocal pairs are clustered they take with them into the higher order types all of the objects with which they had been previously clustered.

Higher order types are clusters of objects associated by dyadic types. If Object  $O_1$  is reciprocal with Object  $O_2$ , which is reciprocal with Object  $O_3$ , which is reciprocal with — — — which is reciprocal with Object  $O_N$ , then all  $N$  objects belong to the same higher-order type. Furthermore, if any Object  $O_i$  in the latter chain is reciprocal with some other Object  $O_1'$ , which is reciprocal with  $O_2'$ , which is reciprocal with  $O_3'$ , which is reciprocal with — — — which is reciprocal with  $O_N'$ , then the  $N + N'$  objects belong to the same higher-order type, and analogously for all other like appendages.

### *Versions of the Method*

Three versions of the method are described and compared empirically: (1) *Concentrated Clustering*, (2) *Dispersed Clustering*, and (3) *Median Clustering*. *Concentrated Clustering* is the recommended version and is outlined first, followed in turn by *Dispersed* and *Median Clustering*.

### *Concentrated Clustering*

The method is described and illustrated simultaneously. The method is applied to a matrix which was difficult to analyze by many methods because many ties in interassociations occur throughout some of the analyses (McQuitty, Price and Clark, 1967), and many of the differences in interassociations are small. This matrix was chosen for two reasons: (1) to demonstrate that the method solves the problem of ties, and (2) to illustrate its effectiveness even when the classifications depend on small differences in interassociations between objects.

The matrix is shown in Table 1. The highest entry in every column is underlined. The second highest entry in some columns is also underlined. This fact can be ignored for the time being.

Each of the highest entries is examined to determine whether or not

it is *reciprocal*. An entry in any column,  $i$ , and row,  $j$ , is reciprocal if, and only if, it is highest in both Columns  $i$  and  $j$ . The highest entry in Column A of Table 1 is 30, and it is with Row P, but the highest entry in Column A is not reciprocal; Object A is not highest in Column P. The highest entry in Column A is for the time being ignored. Analogously, the highest entry in Column B is not reciprocal, and it is also ignored for the time being.

The highest entry in Column C is 34 and occurs in both Rows F and P. It is also highest in both Columns F and P; it is reciprocal between C and F and between C and P. Column C is redesignated  $i_1$  and Column F is redesignated  $j_1$ . Then Column C is redesignated  $i_1'$ , and Column P is redesignated  $j_1'$ .

The next reciprocal pair, from left to right, in the matrix is in Objects D and L. Column D is redesignated  $i_1''$  and Column L is redesignated  $j_1''$ . The only other reciprocal pair in the matrix at this

TABLE I  
Illustrating the Initial Analysis of Concentrated Hierarchical Clustering  
Redesignations

	$i_1'$ $i_1 i_1''$		$j_1$		$i_1''$		$j_1''$		$i_1$		$j_1'$		$i_1'$		$j_1'$		$i_1''$		$j_1''$	
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
A		20	29	20	25	24	20	13	29	20	23	18	28	28	24	30	28	16	22	20
B	20		20	25	17	15	26	20	15	25	13	26	20	20	27	21	25	20	18	25
C	29	20		19	26	34	23	13	33	20	30	18	27	31	26	34	32	19	30	32
D	20	25	19		22	18	25	20	18	22	18	27	20	19	25	19	22	21	19	25
E	25	17	26	22		24	17	23	23	16	28	20	18	23	20	21	24	14	26	14
F	24	15	24	18	24		20	12	22	17	22	15	29	29	23	30	28	16	22	20
G	20	26	23	25	17	20		14	19	30	18	26	18	22	25	23	26	24	16	28
H	13	20	13	20	23	12	14		12	15	16	18	13	11	21	11	13	15	18	12
I	29	15	33	18	23	33	19	12		20	29	15	30	28	26	33	25	17	29	21
J	20	25	20	22	16	17	30	15	20		14	20	21	21	28	22	24	27	15	31
K	23	13	30	18	28	32	18	16	29	14		16	25	24	21	29	22	13	31	16
L	18	26	18	27	20	15	26	18	15	20	16		15	17	23	15	19	21	18	21
M	28	20	27	20	18	29	18	13	30	21	25	15		27	26	30	26	17	23	23
N	28	20	31	19	23	29	22	11	28	21	24	17	27		24	30	27	19	23	22
O	24	22	26	25	20	28	25	21	26	28	21	23	26	24		25	24	22	24	27
P	30	21	24	19	21	30	23	11	22	22	29	15	20	30	25		28	17	26	23
Q	28	25	32	22	24	28	26	13	25	24	22	19	26	27	24	28		18	22	22
R	16	20	19	21	14	15	24	15	17	27	13	21	17	19	22	17	18		14	30
S	22	18	30	19	26	32	16	18	29	15	31	18	23	23	24	26	22	14		16
T	20	25	22	25	14	20	28	12	21	21	16	21	23	22	27	20	22	30	16	
Rank	5	10	12	4	2	13	1	8	15	16	11	14	3	18	7	20	6	17	9	19

Note.—The basic table is from McQuitty, Price and Clark, 1967.

stage is for Objects J and T. Column J is redesignated  $i_1'''$  and Column T is redesignated  $j_1'''$ .

Four clusters, of two objects each, have now been isolated: (1) C F, (2) C P, (3) D L, and (4) J T, as indicated by the redesignations of their columns in the table. The next problem is to decide which member of each cluster better represents the cluster for the next stage of the analysis. That member of a pair which participates in more highest column entries is assumed to be the better representative. It is recognized by the fact that it has more underlines in its row.

If a tie occurs, the selection is made randomly. This was accomplished in this study by assigning randomly the numbers 1 to 20 to the code letters of the objects, A through T, using a table of random numbers. Whenever a tie occurred, the object with the larger code number was chosen. The randomly assigned code numbers are shown in the last row of Table 1.

The number of underlines in Row C was five at this stage of the analysis, as recorded to the left of Row C in the table. Other underlines are added in the course of the analysis, and it is for this reason that Row C has six underlines. The number of underlines in Row F is four. Object C is by assumption a better representative of its cluster than is F. Row F is marked out and eliminated from the rest of the analysis. Analogously, Object C is a better representative than P, and J is a better representative than T. Rows P and T are marked out and eliminated from the rest of the analysis. The members of the other cluster, D and L, are tied with one underline each. Object L was chosen because it has the larger code number, and Object D was eliminated; Row D was marked out. Consistently with having eliminated certain rows, corresponding columns were marked out and eliminated from the rest of the analysis. This completed Step 1.

In Step 2, the highest entries in the columns of the reduced matrix are underlined, as shown in the table. For example, in Column A, the highest entry for the original matrix was 30 in Row P, but Row P was eliminated in Step 1. The highest entry in Column A of the reduced matrix is 29; it appears in Rows C and I. The highest entries in each of the other columns is as shown by the underlines.

Every subsequent step is a repetition of Step 1, except each of them is applied to a matrix reduced by the operations of the previous step. Each step reduces the matrix by one or more corresponding rows and columns, as outlined above in Step 1.

Showing any further analysis in Table 1 would have made the table difficult to read in following the description up to this point. The complete operations of *Concentrated Clustering* are shown in Table 2, using the symbols applied in outlining Step 1. The row of Step 3, for ex-



TABLE 2

[illegible]

**Note**—The basic data is from McQuilty, Price and Clark, 1967.

ample, shows that C (redesignated  $i_3$ ) joined Q (redesignated  $j_3$ ) and that J (redesignated  $i_3'$ ) joined O (redesignated  $j_3'$ ). The column of Step 3 shows that Row C at that stage of the analysis had four underlines and that Q had one underline. Column and Row Q were marked out in Step 3 (as shown by the line through Row Q extending into the Column of Step 3) because Row Q had fewer underlines than Row C. The column of Step 3 shows that J with two underlines tied O, also with two underlines. Consequently, one of the two objects was eliminated randomly (by random numbers in this case) viz., the one with the lower code number. Object J with a code number of 16 was retained and Object O with a code number of seven was eliminated. The rest of the analysis continued in the same fashion as just outlined.

The hierarchical structure which derives from the analysis (Figure 1) can be prepared directly from Table 2, but the operation is simplified and facilitated by an intermediate table, especially so if the matrix is large and diffuse. The intermediate table orders the data according to the size of the reciprocal pairs, which yield the classification, as shown in Table 3. In preparing Table 3, the reciprocal pairs of Table 2 were first ordered numerically from the largest down to the smallest scores

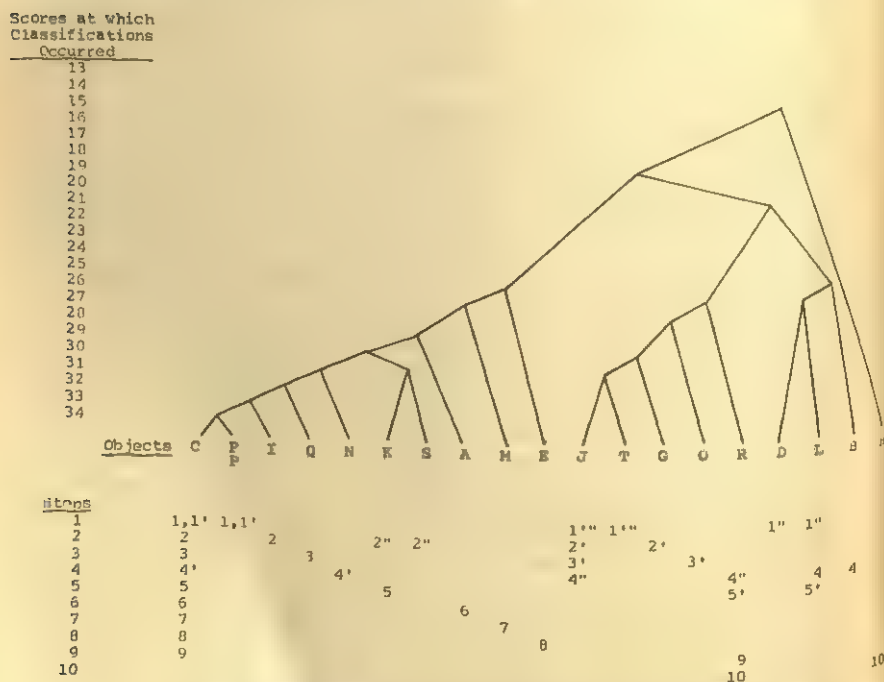


Figure 1. Concentrated Hierarchical Clustering of the Data of Tables 1 and 2.

TABLE 3  
*Scores at Which Classifications Occurred*

Scores	Steps, Reciprocal & Redesignation Pairs
34	1 CF, 1' CP
33	2 CI
32	3 CQ
31	1''' JT, 2'' KS, 4' CN
30	2' GJ, 5 CK
29	6 AC
28	3' JO
27	1'' DL, 4'' JR, 7 CM
26	4 BL, 8 CE
25	
24	
23	
22	
21	5' LR
20	
19	9 CR
18	
17	
16	
15	10 HR

(Column 1). The largest score of a matrix is always involved in the first step; it is reciprocal from the start of the analysis by virtue of the fact that it is the largest score in the matrix and therefore largest in each of the two columns in which it occurs.

Step 1 of Table 2 shows that Object C joined F with the largest score of 34, and under a redesignation of 1 and that C also joined P with the largest score of 34, and under a redesignation of 1'. These two classifications are shown as the first entry of Table 3. Step 1 of Table 2 shows also that Object D joined L with a score of 27 and under a redesignation of 1''. This outcome is shown by 1''DL after 27 in Table 3. Step 1 of Table 2 shows further that Object J joined T with a score of 31 and under a redesignation of 1'''. This outcome is shown by 1'''JT opposite the score of 31 in Table 3. The classification of the other steps of Table 2 were recorded in Table 3 analogously.

The structure of Figure 1 was prepared from Table 3, working from the largest score down and from left to right within every row of Table 3. The top score, 34, of the table shows that C joined F under this score and with a redesignation of 1, and C joined P under this score and with a redesignation of 1'. These three objects are plotted accordingly in Figure 1 with a 1 under each C and F and a 1' under each C and P. Under a score of 33, C joined I with a redesignation of 2. Under a

score of 32, C joined Q with a redesignation of 3. These latter two classifications are shown in Figure 1, with a 2 under each C and I and a 3 under each C and Q. This method of plotting associates the hierarchical structure with the steps of the analysis as reported in Table 2.

Whenever an object of a cluster classifies with another object, as in most of the above examples, it takes with it into the new cluster all of the objects with which it is already classified.

Up to this point the plotting was straight forward and simple. But under a score of 31, J and T joined each other, as did also K and S. None of these objects had already joined the initial cluster, outlined above. It was helpful to start each of them on a separate sheet and to introduce them into the initial cluster (Figure 1) only when they joined it by an association with some member of the initial cluster.

Cluster KS soon joined the initial cluster under a score of 30 between C (of cluster CFPIQN) and K (of Cluster KS). Accordingly, the KS cluster was at this time attached to the initial cluster of Figure 1. However, Cluster JT first expanded on its separate sheet into Cluster JTGOR and then joined a cluster, DLB, which had been started on another separate sheet. These two clusters classified together under a score of 21 between L and R. Cluster DLB was at this time attached to Cluster JTGOR on the sheet on which the latter cluster was initiated. The resultant cluster, JTGORDLB then joined the initial cluster under a score of 19 between C and R and was at this time attached to the initial cluster of Figure 1. The final classification joined Object H with the initial cluster under a score of 15 between H and R, and H was attached to the initial cluster.

Within every cluster, the objects were classified together from left to right as the classification scores decreased. For example, Objects J and T, with the highest score, 31, of the second cluster are shown to the extreme left of their cluster and are then followed from left to right by G, O, and R which joined with scores of 30, 28, and 27 respectively. Analogously, the clusters were arranged from left to right as the size of the scores which attached them to the initial cluster decreased.

### *Dispersed and Median Clustering*

*Dispersed Clustering* differs from *Concentrated Clustering* in only one way. When two objects are reciprocal, *Concentrated Clustering* eliminates from further analysis the one which has the fewer underlines in its row. By way of contrast, *Dispersed Clustering* eliminates the object which has the more underlines in its row. In between these two extremes, *Median Clustering* alternates; the object with the greater number of underlines is eliminated in half of the pairs and the one with the fewer numbers is eliminated in the other half. In the case of a tie,

all three methods eliminate one object of the reciprocal pair randomly.

With a large matrix, *Median Clustering* makes all choices for elimination randomly, because this saves computationally and yields approximately the same number of the two kinds of selections. In the current (small sample) application of *Median Clustering*, the first choice of a non-tied pair was made randomly (selecting the pair the larger code number). This approach selected Object F (with Code 13 and four underlines—Table 1) over Object C (Code 12 and five underlines). This outcome required the next choice for a non-tied pair to be the object with the greater number of underlines. Thereafter, the selection in non-tied pairs alternated in terms of the number of underlines.

Hierarchical structures for *Dispersed* and *Median Clustering* of the data of this study are shown in Figures 2 and 3 respectively.

### *A Comparison of the Results from the Three Versions*

Figures 1, 2 and 3 reveal that *Concentrated Clustering* initiated the fewest clusters by pairs (5, C-F, C-P, K-S, J-T and D-L) and is fol-

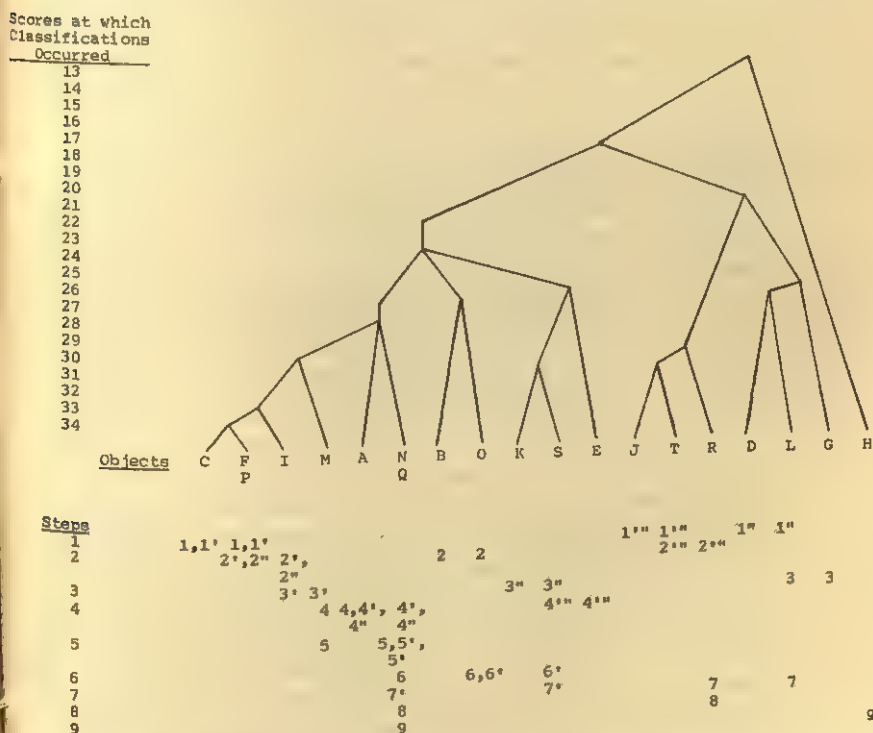


Figure 2. Dispersed Hierarchical Clustering of the Data of Tables 1 and 2.



Scores at which  
Classifications  
Occurred

13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34

Objects

C F I M A K S N Q E J T R G D L B O

Class

1  
2  
3  
4  
5  
6  
7  
8  
9  
10

1,1' 1,1' 2' 3" 3" 3" 3" 4' 5 6 7 8 9 10

Figure 3. Median Hierarchical Clustering of the Data of Tables 1 and 2.

lowed in order by *Median* (6, C-F, C-P, K-S, J-T, D-L and B-O) and *Dispersed* (8, C-F, C-P, A-N, A-Q, B-O, K-S, J-T and D-L).

Entries at the bottom of the hierarchical structure show that *Concentrated*, *Median*, and *Dispersed* versions required 10, 10 and 9 steps respectively to complete the classification of the 20 objects.

Table 4 compares the versions in terms of the scores at which the classifications occurred. All three versions required exactly two classifications at 34, at least one at 31 and at least one at 27. These requirements are due to the fact that all four of these classifications occurred

TABLE 4  
Frequencies and Accumulated Frequencies of Scores at Which Classifications Occurred

Mean	Scores	34	33	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13
27.94	Concentrated	2	1	1	3	2	1	1	3	2	0	0	0	0	1	0	1	0	0	0	1	0	0
		2	3	4	7	9	10	11	14	16	16	16	16	16	17	17	18	18	18	18	19	19	0
27.63	Median	2	1	0	2	2	2	2	2	1	1	1	0	1	0	0	1	1	0	0	0	0	0
		2	3	3	5	7	9	11	13	14	15	16	16	17	17	17	18	19	19	19	19	19	0
27.00	Dispersed	2	1	0	2	2	0	3	2	2	0	2	0	0	1	0	0	1	0	0	0	0	0
		2	3	3	5	7	7	10	12	14	14	16	16	16	17	17	17	18	18	18	18	18	18

in the first step, and all three versions are identical through the first step. Table 4 shows that the *Concentrated* version obtained larger accumulated frequency of classification than the *Median* version for scores 32 through 29 and 27 through 25, equal at 34, 33, 28, 24, 23, 21 through 19 and at 15, and smaller at 22 and 18 through 16. The *Median* version obtained larger accumulated frequencies than the *Dispersed* version for scores 29 through 27, 25, 22 and 19 through 14 and equal at 34 through 30, 26, 24, 23, 21, 20 and 13. The mean score at which the *Concentrated* version classified the 20 objects is 27.94, followed by *Median* with a mean of 27.63 and by *Dispersed* with a mean of 27.00.

All of the above findings are consistent with the fact that the concentrated version retains for further analysis from each reciprocal pair the object which has the more scores highest with other objects (the more underlines in its row). This object has a relatively good chance of entering soon into another reciprocal pair and thus of attaching another object to its cluster (rather than initiating a new cluster). At the same time, the selected object tends to reduce the number of reciprocal pairs which can be realized in the next few subsequent steps; only one of the objects with which it is highest can form a reciprocal pair with it in any one step (without a tie) and by virtue of having it highest with each of them, the associated objects can not form a reciprocal pair with any other object.

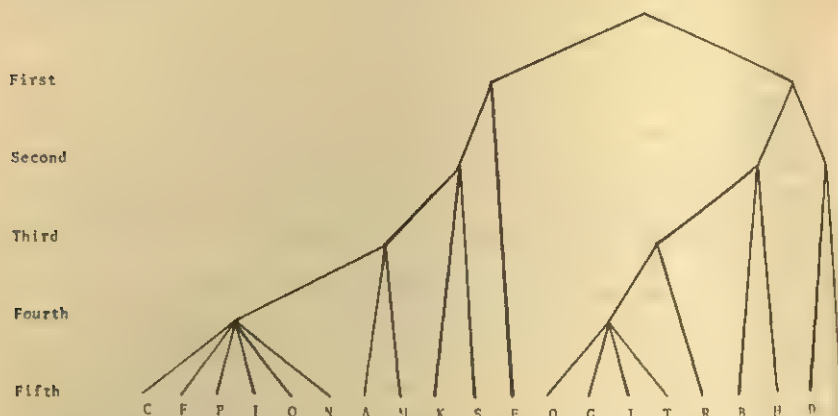
By way of contrast with the above conditions, the other two versions disperse the underlined entries through more rows. This allows for more reciprocal pairs to occur in one or more steps and for the total number of steps to be reduced.

#### *A Comparison with Reliable and Valid Hierarchical Classification*

Reliable and Valid Hierarchical Classifications (McQuitty and Frary, 1971) derives from a more stringent definition of types than applied here. A type is there defined as a category of objects of such a nature that every object in the category is more like every other object in the category than it is like any object in any other category. The method was designed to use "that particular set of indices of association which produces the most reliable and valid solution."

Figure 4 portrays the classification into types by the above method for the data of this study. The method produced two classifications at Level 4 (Clusters CFPIQN and OGJT), two at Level 3, three at Level 2 (CFPIQNAMKS, OGJTRBH, and DL), and two at Level 1. The clusters are arranged from left to right in the order of joining into a

Levels at which  
Classifications  
Occurred\*



\*Levels are labeled from top to bottom because the method partitions from top to bottom.

Figure 4. Reliable and Valid Hierarchical Classifications of the Data of Tables 1 and 2 (McQuitty and Frary, 1971).

larger cluster. The order within an initial cluster is arbitrary; the order within each Cluster CFPIQN and AM for example, is arbitrary.

Table 5 compares the classification results from each version with those from Reliable and Valid Classification. In making these comparisons the arbitrary orders within the initial clusters of the Reliable and Valid method were adjusted to maximize their agreement with each particular version with which they were compared. For example, Figure 4 shows the arbitrary arrangement within an initial cluster to be O, G, J, T, R. The fixed order of these objects in Figure 2, for the *Dispersed* version, is OJTRG, though not in juxtaposition to one another throughout; both O and G are separated from the other three objects. When Reliable and Valid results were compared with those from the *Dispersed* version at the bottom of Table 5, the above five objects were reordered to read OJTRG, but changes from one initial cluster to another, or reordering of clusters were excluded.

The orders of objects having been established (as outlined above) Kendall's tau (Kendall, 1955) was computed for the results from Reliable and Valid with those of each of the three versions of the current paper. The taus are listed in the right hand column of Table 5. They show that the results from the *Concentrated* version is in highest agreement with those from the Reliable and Valid method with a tau

TABLE 5  
*A Comparison of the Three Current Versions with Reliable and Valid*

					Tau
Reliable and Valid	<u>CFPIQN</u>	<u>AM KSE</u>	<u>JTGOR</u>	<u>BH DL</u>	.916
Concentrated	CFPIQN	KS AME	JTGOR	DLB H	
Reliable and Valid	<u>CFPINQ</u>	<u>MA KSE</u>	<u>JTRGO</u>	<u>BH DL</u>	.842
Median	CFPIMA	KSNQE	JTRG DL	BO H	
Reliable and Valid	<u>CFPINQ</u>	<u>MA KSE</u>	<u>OJTRG</u>	<u>BH DL</u>	.80
Dispersed	CFPIM	ANQ BO	KSE JTR	DLG H	

For Reliable and Valid only, the order is arbitrary within each underlined group.

of .916, and they are followed in order by the *Median* and *Dispersed* versions with taus of .842 and .800 respectively. A tau of 0.80 based on 20 cases is significant at value smaller than .00001 for a one-tailed test (Kendall, 1955). These results indicate that all three versions compare reasonably well with the Reliable and Valid method.

### *A Criterion of Internal Consistency*

In developing a criterion of internal consistency, the original matrix is reordered to conform to the hierarchical clustering derived from *Concentrated Clustering*. The entries of the reordered matrix are not, however, the indices of the original matrix. They are instead the ranks of those entries within the columns of the original matrix.

The reordered matrix of the current analysis is shown in Table 6. The objects are ordered from left to right and from top to bottom in the new matrix to conform to their order from left to right in the hierarchical classification of Figure 1. As a consequence C and F are listed first and second from left to right in Table 6 and are followed by P. The entry in each (1) Row C—Column F and (2) Row F—Column C is one because 34 (the agreement score between C and F in Table 1) is highest in each Column C and Column F. The entry in Row P—Column C and Row C—Column P of Table 6 is also 1 because their agreement score, 34, is highest in each of their columns. Thirty-four appears twice in Column C of Table 1 and is the highest entry in the column. There is a tie. In the case of a tie, the rank value which would be assigned if there were only one score at the tied values is assigned to all tied scores.

TABLE 6  
*A Portrayal of the Validity of the Classifications*

	C	F	P	I	Q	N	K	S	A	M	E	J	T	G	O	R	D	L	B	H
C		1	1	1	1	1	3	3	2	5	2	11	9	9	4	8	13	10	9	11
F	1		3	1	2	3	1	1	8	3	5	15	14	12	14	14	17	16	17	15
P	1	5		1	2	2	4	5	1	1	11	7	7	9	7	11	13	16	8	18
I	3	2	2		8	4	4	4	2	1	7	11	12	14	4	11	17	16	17	15
Q	4	8	8	10		6	10	10	4	7	5	6	9	3	10	10	6	9	4	11
N	5	6	3	8	5		8	8	4	5	7	9	9	11	10	8	13	14	9	18
K	6	3	7	5	13	8		2	10	9	1	19	16	15	17	19	17	15	19	7
S	6	3	9	5	13	10	2		11	10	2	17	16	18	10	17	13	10	15	5
A	8	9	3	5	2	4	9	10		4	4	11	14	12	10	14	10	10	9	11
M	9	6	3	4	6	6	7	8	4		14	9	7	15	4	11	10	16	9	11
E	10	9	14	11	10	10	6	5	7	15		16	18	17	19	17	6	7	16	1
J	14	15	13	13	10	14	17	18	12	12	17		1	1	1	2	6	7	4	8
T	13	12	11	12	13	12	14	16	12	10	18	1		2	2	2	5	4	15	
G	12	12	11	14	6	12	12	16	12	15	15	2	3		7	3	2	2	2	10
O	10	11	10	9	10	8	11	7	8	7	12	3	4	6		4	2	4	1	2
R	16	16	17	16	18	16	18	19	18	17	18	4	2	8	16		9	5	9	8
D	16	14	16	15	13	16	12	12	12	13	10	7	5	6	7	5		1	4	3
L	18	17	18	17	17	18	14	13	17	18	12	11	12	3	14	5	1		2	5
B	14	17	14	17	8	15	18	13	12	13	15	5	5	3	2	7	2	2		3
H	19	19	19	19	19	19	14	13	19	19	7	17	19	19	17	16	10	10	9	

The rank for the next score after the tied score is the rank of the tied scores plus the number of them. In this case, the rank of the tied scores is one and the number of them is two to yield a rank of three for the next score, 33, in Column C with Row I of Tables 6 and 1. Ranks for other tied scores were computed in this same fashion.

Any cluster can now be examined in terms of all of its entries to assess how closely it conforms to a type as defined in Reliable and Valid Hierarchical Classification (McQuitty and Frary, 1971). Some of the more distinctive clusters of Figure 1 are emphasized in Table 5 by enclosing them in heavy lines. These were chosen to correspond to the major divisions in clusters as indicated by Figure 1.

If the definition of a type were fulfilled by a cluster, no rank in a cluster would be larger than  $n - 1$ , where  $n$  is the number of objects in the cluster; every object in the cluster would be required by the definition of a type to be more like every other object in the cluster than like any object in any other cluster. Any exception is called a "spot" (McQuitty and Frary, 1971).

There are exceptions to the above requirement. Cluster CFPI-QNKS, for example, contains eight objects. If it conformed to a type as herein defined, it would contain no rank above seven. It contains 14 ranks out of a total of 56 with ranks above seven; it contains 14 "spots." However, the cluster does conform reasonably well to the definition of types.

A difficulty with the outcome is that we do not know whether the



spots are due to fallibility of data or nonconformity of objects to the "above" definition of types. The availability of the methods facilitate investigations into issues of this kind.

### *The Isolation of Stringently Defined Types*

The methods of this paper were designed to isolate types even though they are vaguely delineated in data. The methods can also isolate types when they are clearly delineated in data, as will now be shown.

Let a type be defined as a category of objects which possesses a unique pattern of characteristics; every object of the type possesses all of the characteristics, and no object not in the type possesses all of the characteristics; each of these latter objects possesses instead the pattern of characteristics unique to its type. This definition of types is further restricted to recognize that types higher in a hierarchy (with more members) have fewer characteristics in common.

Let the content on which the objects of an analysis are assessed be so chosen that no matrix or submatrix will yield a reciprocal score for other than members of a type, i.e., wherever this is possible. The only matrix or submatrix in which it is not possible, theoretically, is one which does not contain at least two members of a type.

Let the content on which the objects are assessed be further restricted. Let it be so specific to the different types, that a reciprocal score between nonmembers will be radically lower than all other reciprocal scores; it can then be spotted and ignored.

Under the above conditions, the method will yield clusters in which every member of a cluster is associated with all other members of the cluster through reciprocal pairs. Let  $i, j$ , and  $k$  be any three objects of a higher order type and let them be associated initially by reciprocal pairs  $ij$  and  $jk$ . When  $j$  is classified with  $i$ , it must take with it  $k$  to form the higher order type  $ijk$  (see early section on Initial Definitions of Types), or analogously when  $j$  is classified with  $k$ , it must include  $i$ . At a more complex level,  $j$  must take with it into the higher level type all of the objects with which it had been classified up to that time. By this process all objects are classified into a higher order structure.

### *A General Evaluation of the Method*

Unique values of the method are its ability to analyze rapidly huge matrices of interassociations between objects into statistically defined types and to do this whether or not the types are precisely or loosely delineated by the data.

The problem, as usual in multivariate analyses of this kind, is to

select the proper sample of content on which the objects are assessed so that substantive types can be isolated if they do in fact exist. Trial and error applications of the method to carefully selected content in terms of theoretical departures should facilitate the eventual isolation of types if they do in fact exist.

### *Summary*

This paper develops and illustrates a rapid method for clustering into hierarchical structures large matrices of interassociations between objects, up to  $1000 \times 1000$  and larger, and should facilitate the isolation of substantive types if they do in fact exist.

### REFERENCES

- Kendall, M. G. *Rank correlation methods*. (2nd ed.) New York: Hafner Publishing Company, 1955, p. 5 and 54.
- McQuitty, L. L. Pattern analysis illustrated in classifying patients and normals. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1954, 14, 598-604.
- McQuitty, L. L. A pattern analysis of descriptions of "best" and "poorest" mechanics compared with factor-analytic results. *Psychological Monographs*, 1957, 71 (17, Whole No. 446).
- McQuitty, L. L. and Frary, J. M. Reliable and valid hierarchical classification. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1971, 31, 321-346.
- McQuitty, L. L., Price L., and Clark, J. A. The problem of ties in a pattern analytic method. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1967, 27, 787-796.

## ANALYSIS TECHNIQUES FOR EXPLORATORY USE OF THE MULTITRAIT-MULTIMETHOD MATRIX

MICHAEL L. RAY  
Stanford University

ROGER M. HEELER  
York University

Methods for analyzing the multitrait-multimethod matrix are reviewed and the results of their application to a classic data set are compared. It is shown that different analysis methods can yield different validity conclusions, and that the results obtained are partly dependent on the subjective judgments of the users. It is proposed that several analysis methods should be used in tandem on each data set and their results should be examined for convergence. Multitrait-multimethod matrices should be examined in an exploratory as well as validity testing mode so as not to waste their rich data content.

THE multitrait-multimethod matrix of Campbell and Fiske (1959) has become a standard device for testing the validity of the measures used in educational and psychological research. Like all good innovations, the multitrait-multimethod matrix has spawned a series of criticisms and developments of the original paradigm. One important line of evolution has been the use of the matrix for exploratory research. A series of techniques that aid exploratory use of the matrix have been developed and are reported in this paper, together with a suggestion that the use of several of these techniques in parallel may increase the overall validity of the analysis.

The classic use of the multitrait-multimethod matrix is as a confirmatory technique. The inter-correlations of the several trait-method units are used to test for both the convergent and discriminant validity of the predefined traits, after first establishing that adequate reliability is present. Exploratory analysis of the matrix also tests for the con-

vergent and discriminant validity of traits. But the traits tested may be revised from those prehypoththesized according to the trait-method relationship revealed in the matrix.

The idea of using the rich data content of the multitrait-multimethod matrix in an exploratory mode is not new. Campbell and Fiske (1959, p. 103) wrote, "We believe that a careful examination of a multitrait-multimethod matrix will indicate to the experimenter what his next step should be; it will indicate which methods should be discarded or replaced, which concepts need sharper delineation, and which concepts are poorly measured because of excessive or confounding method variance." More recently Boruch and Wolins (1970), Conger (1971), and Krause (1972) have suggested that the multitrait-multimethod matrix be used in trait development. For example, Krause (1972) suggested that failing a convergent-discriminant validity test is not, in general, disconfirmative of an instrument's validity. Usually failure of a particular criterion is a clue to the strategy of instrument development that is required. Thus no single study is likely to constitute adequate validation of a measure, but each succeeding study should influence both the structure and inference to be derived from the next.

The exploratory approach may be of particular value in association with the extended areas of application to which the original matrix has been applied. The most complete set of suggestions for extended use has been given by Paisley, Collins, and Paisley (1971). They demonstrated how the convergent-discriminant process may be used to validate five elements of research design through the ten matrices formed by all possible pairs of the elements, namely concepts (traits), measures (methods), populations, times, and analysis models. Other authors have made use of one or more of these matrices to meet particular research needs. For example Centra (1971) used a multigroup-multiscale matrix in which different populations replaced the methods of the usual matrix. Werts, Jöreskog, and Linn (1972) used a traits-time matrix formulation as an approach to the analysis of panel data. Fishbein (1967) proposed a multiattitude object—multimethod matrix to test the ability of behavioral measures to discriminate between attitudinal objects or situations. In these varied applications the exploratory development of matrix elements is likely to be at least as important as element confirmation.

Several authors, including Campbell and Fiske (1959), Althausen, Heberlein, and Scott (1971), and Krause (1972) have shown that disparate measurement methods are desirable for a multitrait-multimethod matrix. The use of disparate methods has two desirable properties. First, it increases the likelihood that the methods are un-

correlated. This is essential for the validity of the traditional analysis and is desirable for more recent methods. Second, disparate methods are required if an adequate sampling of the universe of methods is to be obtained. Krause (1972) has noted that an adequate sampling of traits, methods, and populations is required if the results of a multi-trait-multimethod matrix are to be generalizable.

Once again, however, these developments increase the need for exploratory as opposed to confirmatory matrix analysis. It is unlikely that new and disparate methods will produce confirmation without a series of exploratory analyses first being conducted.

### *Analysis Techniques*

Given a need for an exploratory approach, what analysis techniques are available, and what are their strengths and weaknesses? The original Campbell and Fiske analysis approach was a simple examination of the relative sizes of the matrix coefficients. Convergent validation was accomplished by examining the size of the correlations (validities) between different measures of the same trait. A three stage discriminant validation then followed in which the validities were compared with the different-trait, different-method correlations, and the overall pattern of correlations was checked for consistency. This analysis approach provides for a strict confirmatory approach, although validation failures can be used to suggest further measure developments. Althauser, Heberlein, and Scott (1970) have noted some causal path inconsistencies in the Campbell and Fiske analysis. These can be minimized if the methods are orthogonal, the matrix is of greater order than two traits  $\times$  two methods, and if testing effects between measures can be avoided.

A plethora of alternate analysis techniques, mostly mutants of factor analysis, have followed the original Campbell and Fiske (1959) approach. Space limitations prevent review of all these alternatives, but three recent developments will be contrasted: (1) restricted maximum likelihood factor analysis (RMLFA), (2) clustering-nonmetric scaling, and (3) multimethod factor analysis. The first two are described below, and multimethod factor analysis will be covered briefly later in connection with an actual analysis.

The RMLFA technique is a superior factor analytic technique developed by Jöreskog (1969). It allows for the testing of specified models of the underlying trait-method factors, thus making clear what is being tested. Different models may be tested varying in complexity from, for example, an orthogonal traits-factors-only model to a traits-and-methods factors model with correlated factors. The goodness of



fit of each model can be tested by a  $X^2$  statistic, and measure variance broken down into trait, method, and error components. The technique can be used in either confirmatory or exploratory mode, i.e., it can be used to test the prehypoththesized trait-method model, or to examine alternative models.

The technique assumes an underlying linear factor structure to the matrix, and the large sample  $X^2$  statistic used in testing model goodness of fit requires the assumption that the variables are multivariate normal distributed. Correlation coefficients are appropriate matrix entries given the linear formulation and metric variable structure assumed.

The cluster-nonmetric scaling technique (Ray, 1973; Shepard, 1972) has many opposite characteristics to RMLFA. The technique uses cluster analysis to group similar measures, and portrays the clusters in a nonmetric scaling configuration to aid eye interpretation of the groupings. The cluster-nonmetric scaling technique does not permit significance testing of prespecified models, but the measure groupings indicated by the data are clearly shown. In contrast with RMLFA no linearity assumptions are required, and measures of association other than the correlation coefficient can be used.

These divergent characteristics of RMLFA and cluster-nonmetric scaling are advantageous if both techniques are used in parallel on the same data set. It will be shown later that even RMLFA, which appears to leave little room for analyst bias, yields different results for different analysts when applied to a common data set. Even more inconsistencies are to be expected when different analysts each use different analysis methods.

The inconsistencies can be turned to advantage if divergent analysis methods are compared in a convergent validation of analysis methods. Results common to two or more disparate techniques will be more convincing than those which appear in one technique only.

### *An Illustration*

A good matrix for comparison of alternative multitrait-multimethod matrix analyses is given by Campbell and Fiske (1959, p. 96, Table 12). It contains the intercorrelations for five traits: "assertive," "cheerful," "serious," "unshakeable poise," and "broad interests" measured by the three methods of "staff ratings," teammate ratings," and "self-ratings," for a group of clinical psychologists.

The original analysis of this clinical psychologist matrix by Campbell and Fiske (1959) found the trait "assertive" to have both convergent and discriminant validity. The three traits "cheerful,"

"serious," and "broad interests" achieved convergent validity and substantial discriminant validity. The trait "unshakeable poise" achieved neither convergent nor discriminant support.

This matrix was subsequently analyzed by the third analysis development mentioned above: Jackson's (1969) multimethod factor analysis.

For this approach Jackson converts the monomethod triangles (correlations between traits as measured by the same method) into identity matrices and factor analyzes the resultant modified multitrait-multimethod matrix. When multi-method factor analysis was applied to the clinical psychologist matrix, all five traits were found valid. It should be noted, however, that multimethod factor analysis is prone to suggesting strong evidence of validity even where other analyses methods indicate some traits with weak validity. Further, Jackson's approach is limited by the initial conversion of the mono-method triangles. In the current example, this results in 30% of the matrix entries being discarded, which is a substantial information loss.

The RMLFA technique has been applied to the clinical psychologist data by two sets of authors, Boruch and Wolins (1970) and Jöreskog (1971).

Jöreskog (1971) first fitted a model consisting of five trait factors only. The  $X^2$  goodness of fit statistic yielded by the technique indicated that this model was not a good fit to the data, i.e., the  $X^2$  of 140.46 was too large for the available 80 degrees of freedom of the model. Jöreskog's next model added three method factors to the five trait factors and yielded an acceptable fit. However, the factor intercorrelations given by the technique indicated that the "staff" and "self" methods factors were unit correlated, so these two factors were combined into one with an acceptable  $X^2$  of 61.51 with 64 degrees of freedom. This final solution contained five trait and two method factors. The factor loadings of each measure on each trait were given, together with the factor intercorrelations and the trait-method-error variance components of each measure. The factor intercorrelations showed that the trait "cheerful" had fairly high correlations with "assertive" and "unshakeable poise" so Jöreskog observed that it was probably confused with these traits. He might also have noted the fairly high negative correlation between "cheerful" and "serious." All methods yielded good measures of "assertive." The method "staff ratings" was best for this trait with a trait variance of .76, method variance of .01, and an error variance of .39. "Cheerfulness" and "unshakeable poise" were also best measured by "staff ratings," but "serious" was best measured by "teammate ratings," and "broad interests" was best measured by "self-ratings."

Boruch and Wolins (1970) used an initial 10 factor model consisting of five traits, three methods, and one general factor. This was reduced to an eight factor model by combining the traits "assertive" and "cheerful" and the methods "staff" and "self." "Cheerful" was judged not to be a distinctive trait because it loaded heavily on the general factor and had low loadings on the combined "assertive-cheerful" factor. The other four traits were found to be valid because of their low loadings on the general factor and high loadings on specific factors. "Unshakeable poise" was well measured only by "only staff ratings."

It is interesting to compare the analytic solutions obtained by Jöreskog and Boruch and Wolins. Both used the same general analysis technique and an exploratory mode, but obtained different, well fitting solutions. Boruch and Wolins included a general factor; Jöreskog did not. Both combined the methods "staff ratings" and "self ratings"; but while Jöreskog maintained "cheerful" as a separate trait, albeit linked with "assertive" and "unshakeable poise," Boruch and Wolins found no valid measure of "cheerful" and combined it with "assertive." These differences illustrate that RMLFA, despite its test statistic, is a technique that in part is dependent on the subjective judgments of its users.

It is also interesting to compare the RMLFA solutions with the Campbell and Fiske (1959) and Jackson (1969) solutions. Campbell and Fiske found the trait "unshakeable poise" invalid but the other four traits have satisfactory validity. This is in reasonable agreement with Jöreskog (1971) solution, because whilst Jöreskog found "unshakeable poise" valid, he observed that it could only be satisfactorily measured by "staff ratings." Campbell and Fiske noted an agreement between "staff ratings" and "teammate ratings," in contrast to the grouping of "staff ratings" and "self ratings" found by the RMLFA analyses. Boruch and Wolins (1970) analysis also differs from Campbell and Fiske's in the treatment of traits. "Cheerful" was valid for Campbell and Fiske but not for Boruch and Wolins, whilst the reverse was true for "unshakeable poise." The Jackson (1969) results differed from all other analyses in finding substantial validity for all traits. These differences illustrate that the procedures for analyzing multitrait-multimethod matrices do not necessarily yield identical results.

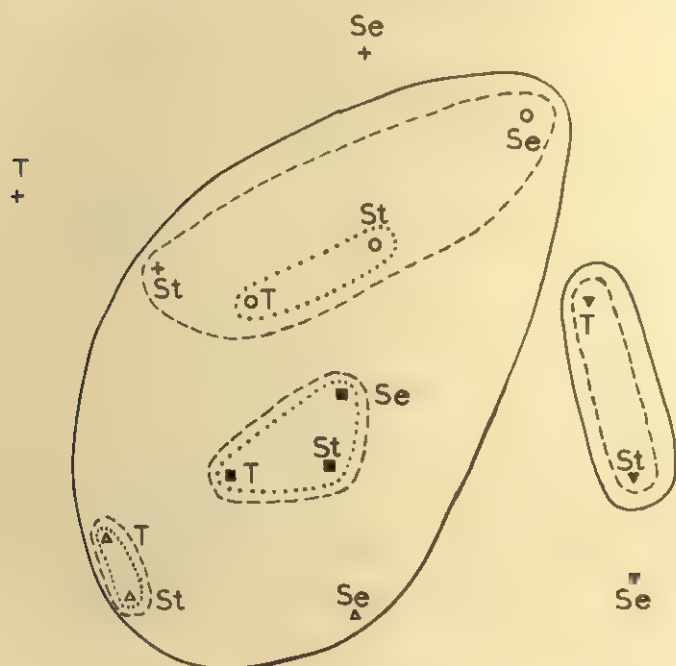
A fifth analysis of the clinical psychologist data was done for this paper. This time the clustering-nonmetric scaling approach was used, and the results parallel those of Campbell and Fiske.

Table I shows the clusters obtained with Johnson's (1967) hierarchical clustering. The clusters are shown both in terms of order of formation and in Gruvaeus and Wainer's (1972) uniquely ordered



format. Figure 1 shows the same cluster analysis, at three levels of clustering, within a two dimensional Euclidean measure space obtained from a nonmetric scaling (Kruskal, 1964) of the data. This form of display aids perception of the clustering (Shepard, 1972).

The results indicate that this is a matrix containing considerable trait validity, since the clusters mostly form around traits rather than around methods. "Assertive" appears to be the strongest trait, while "cheerful," "broad interests," and "serious" also form distinctive clusters, as shown in Figure 1. "Unshakeable poise" never forms a dis-



### TRAITS

- = Assertive
- = Cheerful
- ▼ = Serious
- + = Unshakeable Poise
- ▲ = Broad Interests

### METHODS

- St = Staff Ratings
- T = Teammate Ratings
- Se = Self Ratings

### CLUSTERING LEVELS

- =  $r = .37$
- - - =  $r = .42$
- ..... =  $r = .47$

Figure 1. Two-dimensional MDSCAL configuration of the clinical assessment data, showing three levels of HICLUS clustering. The "stress" (Kruskal, 1964) of this figure was .20. The analysis was checked with a three-dimensional figure having a stress of .11.



tinctive cluster, but appears to be associated with "cheerful." The methods "staff rating" and "teammate rating" appear to be closely associated. They cluster together first in each of the four traits which form distinctive clusters. The visual display from the nonmetric scaling further emphasized the pure cluster analysis findings. Those trait measures which are not in individual clusters with other measures of the same trait, are close to those other same trait measures. The "cheerful" and "unshakeable poise" measures are located in the same spatial zone.

Like Campbell and Fiske (1959), the cluster results indicate "assertive" to be the strongest trait, "cheerful," "serious," and "broad interests" to have good validity, and "unshakeable poise" to have minimal validity. Like Campbell and Fiske, the cluster analysis also found an association between "staff ratings" and "teammate ratings."

The cluster results differed from both RMLFA analyses in this associating of methods. Both RMLFA analyses associated "staff ratings" and "self ratings." The cluster trait results were similar to Jöreskog's results, but differed from Boruch and Wolins (1970) results. "Cheerful" was valid for the cluster analysis but not for Boruch and Wolins, whilst the reverse was true for "unshakeable poise."

If the four techniques and five analyses are considered together it can be seen that measures for the traits "assertive," "serious," and "broad interests" are supported by all. Measures for "cheerful" are supported by Campbell and Fiske (1959), Jöreskog (1971), and cluster analysis, but not by Boruch and Wolins (1970). Thus the preponderance of evidence seems to support "cheerful." "Unshakeable poise" was supported only by the RMLFA and multithreshold factor analysis results. Jöreskog found only the "staff ratings" method of measuring "unshakeable poise" satisfactory. A trait with only one measure may be distinct, but its validity is hardly proven in a convergent sense. Thus "unshakeable poise" appears to be doubtfully measured. There is evidence in both the cluster and Jöreskog analyses that "unshakeable poise" is associated with other traits.

In all five analyses described above, correlation coefficients were used as the information statistics in the matrix. Krause (1972) notes that measurement sets can be codimensional but have low correlation. Thus correlation coefficients under-represent the codimensionality present, thereby decreasing the frequency of convergent validation and increasing the frequency of discriminant validation. Krause (1972) suggests that other measures, such as coefficients of ordinal consistency, should be used if possible.

It is a strength of the cluster-nonmetric scaling analysis that varied

information coefficients can be used without contravening the technique's data requirements. In general the corroborative power obtained by "convergence of analysis methods" will be increased if different information coefficients are used in addition to different analysis methods. The cluster-nonmetric scaling analysis method provides this opportunity. Unfortunately no other information coefficients were available for the clinical psychologists matrix. However, Goodman and Kruskal's Gamma coefficient has been used by Ray (1973) in a multiple measure study of political attitudes. The political study used five measures of three traits measured across precinct groups. A dismal picture of weak convergent validation and non-existent discriminant validation was found for the traits in their original formulation, but the measure clusters suggested several plausible new trait formulations that could be of value in further research.

### Conclusion

Only rarely will a single multitrait-multimethod matrix study suffice for validation. More generally a series of studies will hone the limits of generalizability of a trait-method combination. In these circumstances an exploratory analysis approach is desirable, so that information use from each succeeding study is maximal. Several exploratory analysis techniques have been described. If several of these techniques, using varied information coefficients, are used in parallel, a valuable restraint is provided on the judgmental bias effects to which each individual technique is subject. This does not replace testing with new data sets and the Monte Carlo evaluation of the several techniques, but it does increase the likelihood of valid analysis.

### REFERENCES

- Althauser, R. P., Heberlein, T. A., and Scott, R. A. A causal assessment of validity: The augmented multitrait-multimethod matrix. In H. J. Blalock, Jr. (Ed.), *Causal models in the social sciences*. Chicago: Aldine-Atherton, 1971, 374-399.
- Boruch, R. F. and Wolins, L. A procedure for estimation of trait, method, and error variance attributable to a measure. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1970, 30, 547-574.
- Campbell, D. T. and Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, 81-105.
- Centra, J. A. Validation by the multigroup-multiscale matrix: An adaptation of Campbell and Fiske's convergent and discriminant validation procedure. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1971, 31, 675-683.

- Conger, A. J. Evaluation of multimethod factor analysis. *Psychological Bulletin*, 1971, 75, 416-420.
- Fishbein, M. Attitude and the prediction of behavior. In M. Fishbein (Ed.), *Readings in attitude theory and measurement*. New York: Wiley, 1967, 477-492.
- Fiske, D. W. Consistency of the factorial structures of personality ratings from different sources. *Journal of Abnormal and Social Psychology*, 1949, 44, 218-224.
- Gruvaeus, G. and Wainer, H. Two additions to hierarchical cluster analysis. *British Journal of Mathematical and Statistical Psychology*, 1972, 25, 200-206.
- Jackson, D. N. Multimethod factor analysis in the evaluation of convergent and discriminant validity. *Psychological Bulletin*, 1969, 72, 30-49.
- Johnson, S. C. Hierarchical clustering schemes. *Psychometrika*, 1967, 32, 241-254.
- Jöreskog, K. G. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 1969, 34, 183-202.
- Jöreskog, K. G. Statistical analysis of sets of congeneric tests. *Psychometrika*, 1971, 36, 109-133.
- Krause, M. S. The implications of convergent and discriminant validity data for instrument validation. *Psychometrika*, 1972, 37, 179-186.
- Kruskal, J. B. Nonmetric multidimensional scaling: A numerical example. *Psychometrika*, 1964, 29, 115-129.
- Paisley, M. B., Collins, W. A., and Paisley, W. J. The convergent-discriminant matrix: Multitrait-multimethod logic extended to other research decisions. Unpublished paper. Stanford University, 1971.
- Ray, M. L. Final analysis of a multiple and unobtrusive field study of attitudes and behavior. Report on NIMH grant MH-17869-1 and NSF grant GS 2683, Stanford University, 1973.
- Shepard, R. N. A taxonomy of some principal types of data and of multi-dimensional methods for their analysis. In R. N. Shepard, A. K. Romney, and S. B. Nerlove, *Multidimensional Scaling*, Vol. 1. New York: Seminar, 1972, 21-47.
- Webb, E. J., Campbell, D. T., Schwartz, R. D., and Sechrest, L. *Unobtrusive measures: Nonreactive research in the social sciences*. Chicago: Rand McNally, 1966.
- Wers, C. E., Jöreskog, K. G., and Linn, R. L. A multitrait-multimethod model for studying growth. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1972, 32, 655-678.



## BEHAVIOR OF THE PRODUCT-MOMENT CORRELATION COEFFICIENT WHEN TWO HETEROGENEOUS SUBGROUPS ARE POOLED<sup>1</sup>

ALAN L. SOCKLOFF

Temple University

An equation was derived to determine the relationship between the pooled within-subgroup correlation coefficient,  $r_w$ , and the correlation coefficient obtained from the total group data,  $r_t$ . It was, thus, possible to assess the amount of distortion introduced by pooling heterogeneous subgroups. As a basis for deciding whether to pool two subgroups in order to calculate  $r_t$ , a two-stage procedure was recommended: (1) comparison of the two within-subgroup  $r$ 's; and (2) comparison of  $r_t$  and  $r_w$ . On the basis of results for the second stage test, distortion in  $r_t$  was shown to be a function of the pattern of subgroup mean differences, total group sample size, and the magnitude of  $r_w$ . Implications were discussed.

FREQUENTLY, in the psychological and educational literature, correlational studies are reported in which product-moment correlation coefficients are calculated between two variables for sets of data pooled across two or more, possibly heterogeneous, subgroups. The effect of pooling heterogeneous subgroups to calculate a product-moment correlation coefficient was first noted by Pearson (1896, p. 283) and later discussed by Pearson, Lee, and Bramley-Moore (1899, pp. 274-278). By way of illustration, Pearson et al. presented correlational data between length and breadth of skulls for 805 males ( $r = .0869$ ) and 340 females ( $r = -.0424$ ). When the two subgroups were pooled, an  $r$  of .1968 was obtained, and this  $r$  was considered to represent a large spurious value. These authors concluded: "The correla-

---

<sup>1</sup> Portions of this study were presented at the American Educational Research Association convention, Chicago, 1974.



tion may properly be called spurious, yet as it is almost impossible to guarantee the absolute homogeneity of any community, our results for correlation are always liable to an error, the amount of which cannot be foretold (p. 278)."

Awareness of the problem was carried through the various revisions of Yule's text as a short section in a chapter covering miscellaneous theorems related to correlations. In the 11th edition, Yule and Kendall (1937) offered the following theorem: *"If X and Y uncorrelated in each of two records, they will nevertheless exhibit some correlation when the two records are mingled, unless the mean value of X in the second record is identical with that in the first record, or the mean value of Y in the second record is identical with that in the first record, or both (p. 301)."*

From a sampling of recent introductory statistics textbooks in psychology and education, it was found that writers do discuss the effect on the correlation coefficient resulting from pooling heterogeneous subgroups (Games and Klare, 1967; Glass and Stanley, 1970; Guilford, 1965; Walker and Lev, 1969; among others). Where references are made, Dunlap's (1937) paper on the combinative properties of correlation coefficients and Lindquist's (1940, pp. 219-228) text are most frequently cited. Dunlap, admitting to summarizing a "well-known" method, derived an equation for calculating a total group correlation coefficient from within-subgroup correlation coefficients, means, and standard deviations. Lindquist presented a practical solution to the problem of pooling heterogeneous subgroups by distinguishing the total group correlation coefficient from the pooled within-subgroup correlation coefficient, noting that the latter may be interpreted as an average of the subgroup correlation coefficients. Under the assumption of homogeneous correlation across subgroups, Lindquist argued that the pooled within-subgroup correlation coefficient is the preferred measure insofar as it is more stable across randomly drawn subgroups.

To date, a mathematical formulation of the effect on the correlation coefficient from pooling heterogeneous subgroups has been lacking. In lieu of such a formulation, textbook writers have tended to stress cautious interpretation and the use of subgroup correlation coefficients to help provide a rational explanation for correlational results in the total group data. The major interest of this paper is the derivation of a mathematical formulation and the demonstration of the effects of pooling two subgroups on the total group correlation coefficient. Of additional interest is a procedure to guide the decision concerning the pooling of data for the purpose of calculating a single correlation coefficient.

*Formulation*

Given two subgroups, let  $n_1$  and  $n_2$  be the subgroup sample sizes, where  $n_1 + n_2 = N$ . Let  $U$  and  $V$  be the distances between the subgroup means for  $X$  and  $Y$ , respectively, i.e.,  $U = \bar{X}_2 - \bar{X}_1$  and  $V = \bar{Y}_2 - \bar{Y}_1$ .

*Sum of Squares and Cross Products*

The sum of cross products for the total group,  $SS(XY)_t$ , is defined

$$SS(XY)_t = SS(XY)_1 + SS(XY)_2 + UV\left(\frac{n_1 n_2}{N}\right), \quad (1)$$

where  $SS(XY)_1$  and  $SS(XY)_2$  are the within-subgroup sums of cross products. Since the sum of squares for  $X$  in the total group is actually the sum of cross products with respect to itself,

$$SS(X)_t = SS(X)_1 + SS(X)_2 + U^2\left(\frac{n_1 n_2}{N}\right). \quad (2)$$

Similarly, for  $Y$ ,

$$SS(Y)_t = SS(Y)_1 + SS(Y)_2 + V^2\left(\frac{n_1 n_2}{N}\right). \quad (3)$$

*Total Group  $r_t$* 

Using Equations 1, 2, and 3, the correlation coefficient for the total group is:

$$r_t = \frac{SS(XY)_1 + SS(XY)_2 + UV(n_1 n_2 / N)}{\sqrt{\{SS(X)_1 + SS(X)_2 + U^2(n_1 n_2 / N)\} \{SS(Y)_1 + SS(Y)_2 + V^2(n_1 n_2 / N)\}}}$$

An equivalent form of this equation was derived by Dunlap (1937). If  $SS(XY)_w = SS(XY)_1 + SS(XY)_2$ ,  $SS(X)_w = SS(X)_1 + SS(X)_2$ , and  $SS(Y)_w = SS(Y)_1 + SS(Y)_2$ , then the equation defining  $r_t$  may be simplified:

$$r_t = \frac{SS(XY)_w + UV(n_1 n_2 / N)}{\sqrt{\{SS(X)_w + U^2(n_1 n_2 / N)\} \{SS(Y)_w + V^2(n_1 n_2 / N)\}}}. \quad (4)$$

The correlation coefficient for the total group is, therefore, expressed in terms of pooled within-subgroup sums of squares and cross products, subgroup sample sizes, and distances between the subgroup means.

*Pooled Within-Subgroup  $r_w$* 

The pooled within-subgroup correlation coefficient is obtained from pooling of the subgroup sums of squares and cross products:

$$r_w = \frac{SS(XY)_1 + SS(XY)_2}{\sqrt{\{SS(X)_1 + SS(X)_2\} \{SS(Y)_1 + SS(Y)_2\}}} \\ = \frac{SS(XY)_w}{\sqrt{SS(X)_w SS(Y)_w}} \quad (5)$$

It should be clear from inspection of Equation 5 that for equal variances of  $X$  in both subgroups and equal variances of  $Y$  in both subgroups,  $r_w$  is a weighted arithmetic mean of the within-subgroup correlation coefficients, weighted by the number of observations in each subgroup.

Furthermore,  $r_w$  may be compared to  $r_t$  in two ways. First,  $r_w$  is a special case of  $r_t$  resulting when subgroup mean differences are nonexistent (i.e.,  $U = V = 0$ ). Second,  $r_w$  is that special case of  $r_t$  when subgroup differences are eliminated statistically. The latter comparison requires the form of a first-order partial correlation  $r_{x_t y_t \cdot z}$ , where  $X_t$  and  $Y_t$  are the two variables measured in the total group and  $Z$  is a dichotomous variable indicating subgroup membership. In the formula for the first-order partial correlation coefficient, since  $r_{x_t z}$  and  $r_{y_t z}$  are point-biserial correlation coefficients, the result of operating upon this formula is:

$$r_{x_t y_t \cdot z} = \frac{SS(XY)_t - UV(n_1 n_2 / N)}{\sqrt{\{SS(X)_t - U^2(n_1 n_2 / N)\} \{SS(Y)_t - V^2(n_1 n_2 / N)\}}}$$

Since the above equation represents an alternative definition of  $r_w$ , then  $r_w$ , the pooled within-subgroup correlation coefficient, is also the result of statistically eliminating subgroup differences from the total group correlation coefficient. The latter definition of  $r_w$  suggests that  $r_w$  can be meaningfully used as a descriptive statistic with a known sampling distribution.

*Further Derivation of  $r_t$* 

If the numerator and denominator of Equation 4 are each divided by the product of the pooled within-subgroup standard errors of the mean ( $s_{\bar{x}_w}$  and  $s_{\bar{y}_w}$ ), a more convenient definition of  $r_t$  arises. To complete this series of operations, by defining  $t_x = U/s_{\bar{x}_w}$  and  $t_y = V/s_{\bar{y}_w}$  as subgroup mean differences measured in units of standard errors, the following final form results:

$$r_t = \frac{(N-2)r_w + t_x t_y}{\sqrt{\{(N-2) + t_x^2\}\{(N-2) + t_y^2\}}}. \quad (6)$$

In this form,  $r_t$  is defined in terms of total group sample size, the pooled within-subgroup correlation coefficient, and subgroup mean differences that are measured in units distributed as Student's  $t$  under the assumptions underlying  $t$  tests on means.

### *The Decision to Pool*

The following is a simple, approximate, two-stage procedure, recommended as a basis for the decision to pool two subgroups in order to calculate a single correlation coefficient for the total group data. Underlying both stages is the assumption that the two subgroups were sampled from the same bivariate normal population.

1. The two within-subgroup correlation coefficients should be compared. Under the hypothesis  $H_0: \rho_1 = \rho_2$ , the unit-normal  $z$  test, relying on Fisher's  $r$ - $z$  transformation ( $z_f$ ), can be used:

$$z = \frac{z_{f_1} - z_{f_2}}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \quad (7)$$

If  $H_0$  is rejected, it is unreasonable to calculate  $r_w$  as a measure of correlation for the two subgroups. (For the skull data of Pearson et al. (1899),  $H_0$  would have been rejected ( $z = 1.99, p < .05$ ), and neither  $r_w$  nor  $r_t$  would have been calculated.) If  $H_0$  is not rejected, then  $r_w$  can be considered a useful measure of correlation for the two subgroups, and the second stage test should be followed.

2. In order to assess the distortion introduced by pooling the two subgroups,  $r_t$  should be compared to  $r_w$  under the hypothesis  $H_0: \rho_t = \rho_w$ . For this test, if  $r_w$  is considered an asymptotic estimate of  $\rho_w$ ,

$$z = \frac{z_{f_t} - z_{f_w}}{\sqrt{\frac{1}{N - 4}}} \quad (8)$$

If  $H_0$  is rejected, then  $r_t$  may be considered distorted, but  $r_w$  can be used as a measure of correlation for the two subgroups. If  $H_0$  is not rejected, then pooling of the data from the two subgroups in order to calculate a total group correlation coefficient appears to be a parsimonious and reasonable procedure.

### *Examples of Distortion*

According to Equation 6, distortion in  $r_t$  is affected by subgroup centroid differences and total group sample size. Three patterns of

subgroup centroid differences are of interest. (1) If the mean of Subgroup 2 is higher than the mean of Subgroup 1 on both variables, the greater the difference between the subgroups on the two variables, the more exaggerated the value of  $r_t$  in a positive direction. (2) If the mean of Subgroup 2 is higher than the mean of Subgroup 1 on one variable, and equal on the other variable, the greater the difference between the two subgroups on the one variable, the closer  $r_t$  is to the value zero. (3) If the mean of Subgroup 2 is higher than the mean of Subgroup 1 on one variable, and lower on the other variable, the greater the difference between the subgroups on the two variables, the more exaggerated the value of  $r_t$  in a negative direction. Furthermore, for constant differences between the subgroup centroids as measured in standard errors, increasing the total group sample size serves to minimize the effects of subgroup centroid differences, i.e.,  $r_t$  approaches  $r_w$ .

The effects of subgroup sample size discrepancy on the calculation of  $r_t$  and  $r_w$  can be shown by reference to Equations 4 and 5. According to Equation 5, for constant total group sample size, the larger the discrepancy between the subgroup sample sizes, the greater the influence of the larger subgroup in the calculation of  $r_w$ . In addition, according to Equation 4, for constant differences between subgroup centroids and for constant total group sample size, the larger the discrepancy between  $n_1$  and  $n_2$ , the smaller the effect of subgroup centroid differences in the calculation of  $r_t$ , and, thus, the more equal the values of  $r_t$  and  $r_w$ .

For the second stage test, in order to demonstrate the amount of distortion introduced by pooling heterogeneous subgroups, Equation 6 was employed to calculate  $r_t$  under varying sample conditions. The sample conditions were derived from combinations of four magnitudes of subgroup centroid differences for the three patterns, five total group sample sizes ( $N$ ), and three values of  $r_w$ . The magnitude of difference between subgroup means can be represented by employing four-decimal critical values of Student's  $t$  distribution for  $p < .05$ ,  $p < .01$ ,  $p < .001$ , and  $p < .0001$ , obtained for  $N - 2$   $df$  from Sockloff and Edney's (1972) tables. The five total group sample sizes were 10, 50, 100, 200, and 1000. The three values of  $r_w$  were .8000, .4000, and 0.0000, chosen to represent high, moderate, and low correlations, respectively.

Tables 1, 2, and 3 present calculated values of  $r_t$  derived from the values of  $r_w$  under the varying sample conditions. Also included in these tables are the second stage two-tailed tests of  $H_0: \rho_t = \rho_w$  to assess the amount of distortion introduced under the conditions. Negative values of  $r_w$  are not shown in the tables since the effects of the



TABLE 1

Values of  $r_t$  Resulting from Three Patterns of Subgroup Centroid Differences and Five Total Group Sample Sizes:  $r_w = .8000$

Total group sample size	Significance levels for $t$ distribution when subgroup mean differences equal critical values			
	$p < .05$	$p < .01$	$p < .001$	$p < .0001$
Pattern 1: $\bar{X}_2 > \bar{X}_1$ , $\bar{Y}_2 > \bar{Y}_1$ , both significant				
10	.8799	.9169	.9521	.9727*
50	.8155	.8261	.8408	.8546
100	.8077	.8132	.8210	.8288
200	.8039	.8066	.8107	.8148
1000	.8008	.8013	.8022	.8030
Pattern 2: $\bar{X}_2 > \bar{X}_1$ , significant; $\bar{Y}_1 = \bar{Y}_2$				
10	.6200	.5156	.3914	.2953
50	.7683	.7460	.7138	.6822
100	.7844	.7732	.7568	.7403
200	.7923	.7867	.7784	.7699
1000	.7985	.7973	.7957	.7940
Pattern 3: $\bar{X}_2 > \bar{X}_1$ , $\bar{Y}_2 < \bar{Y}_1$ , both significant				
10	.0813*	-.2523***	-.5691****	-.7547****
50	.6602*	.5654**	.4332****	.3089****
100	.7305	.6816**	.6108***	.5412****
200	.7653	.7405*	.7040**	.6672****
1000	.7931	.7881	.7806	.7729*

Note.—Asterisks refer to significance levels of unit-normal  $z$  tests comparing  $r_t$  and  $r_w$ .

\*  $p < .05$ .

\*\*  $p < .01$ .

\*\*\*  $p < .001$ .

\*\*\*\*  $p < .0001$ .

three patterns for negative  $r_w$ 's are opposite in sign from those shown for positive  $r_w$ 's, e.g., the amount of exaggeration in a positive direction for a positive correlation under Pattern 1 is equal to the amount of exaggeration in a negative direction for a negative correlation under Pattern 3.

As shown in Table 1, under Patterns 1 and 2, large differences between the subgroup means have a small effect on the value of  $r_t$  when  $r_w = .8000$ . Under both patterns, a total group sample size of 50 appears to be sufficient to minimize the distortion introduced by subgroup mean differences that are significant at the .0001 level. On the other hand, the results were quite different under Pattern 3. When the means of the two subgroups are significantly different at the .0001 level, but in opposite directions, for a total group sample size of 50  $r_t$  was calculated to be .3089, which is significantly different from an  $r_w$  of .8000. Furthermore, even for a total group sample size of 1000, the pooling of subgroups when subgroup means differ in opposite directions at the .0001 level produced an  $r_t$  of .7729. Although this value of

$r_t$  is significantly different from  $r_w$  at the .05 level, one can argue that such statistically significant differences between  $r_t$  and  $r_w$  have little practical significance.

According to Table 2, when  $r_w = .4000$ , the results for Pattern 1 suggest that total group sample sizes of 50 are sufficient to avoid distortions introduced by pooling subgroups when both sets of subgroup means differ significantly in the same direction at the .0001 level. Under Pattern 2, significant distortion was not found, even for a total group sample size of 10. The Pattern 3 results suggest that a total group sample size of 200 will avoid distortion in the calculation of  $r_t$ .

According to Table 3, total group sample sizes of 50 appear to be sufficient to avoid distortion when  $r_w = 0.0000$  and the two sets of subgroup means differ at the .0001 level. Based on the symmetry of the sampling distributions of  $r$  when  $\rho = 0$ , this conclusion holds for subgroup means differing in the same or opposite directions.

TABLE 2  
Values of  $r_t$  Resulting from Three Patterns of Subgroup Centroid Differences and Five Total Group Sample Sizes:  $r_w = .4000$

Total group sample size	Significance levels for $t$ distribution when subgroup mean differences equal critical values			
	$p < .05$	$p < .01$	$p < .001$	$p < .0001$
Pattern 1: $\bar{X}_2 > \bar{X}_1, \bar{Y}_2 > \bar{Y}_1$ , both significant				
10	.6396	.7508	.8564	.9182**
50	.4466	.4782	.5223	.5637
100	.4232	.4395	.4631	.4863
200	.4116	.4198	.4320	.4443
1000	.4023	.4040	.4065	.4090
Pattern 2: $\bar{X}_2 > \bar{X}_1$ , significant; $\bar{Y}_1 = \bar{Y}_2$				
10	.3100	.2578	.1957	.1477
50	.3842	.3730	.3569	.3411
100	.3922	.3866	.3784	.3701
200	.3961	.3933	.3892	.3850
1000	.3992	.3987	.3978	.3970
Pattern 3: $\bar{X}_2 > \bar{X}_1, \bar{Y}_2 < \bar{Y}_1$ , both significant				
10	-.1590	-.4184*	-.6648**	-.8092***
50	.2913	.2175	.1147*	.0181**
100	.3459	.3079	.2529	.1987*
200	.3730	.3537	.3253	.2967
1000	.3946	.3907	.3849	.3789

Note.—Asterisks refer to significance levels of unit-normal  $z$  tests comparing  $r_t$  and  $r_w$ .

\*  $p < .05$ .

\*\*  $p < .01$ .

\*\*\*  $p < .001$ .

## Discussion

The various results clearly suggest the varieties of distortion that may be introduced by *haphazardly* pooling subgroups of data for the purpose of calculating a single correlation coefficient. The two-stage test procedure should offer protection against such distortion. In addition, it was shown that greater latitude exists in terms of non-distorting pooling when the subgroup mean differences are small, the subgroup sample sizes are large, and the pooled within-subgroup correlations are low to moderate. The calculated examples suggest limits within which distortion does not seriously affect correlational results.

The types of subgroups to which this discussion refers are those resulting from natural dichotomies and those resulting from an arbitrary split where (a) the decision to split the total group was based on considerations other than that of ridding the data of nonlinearity, and (b) middle range data has been discarded. When an arbitrary split is made to rid the total group data of nonlinearity, the two subgroups may also show evidence of different, but linear, relationships. According to the first stage test, if the two within-subgroup correlations are different, then it would appear unreasonable to even consider pooling the data on the basis of the original rationale for having made the split. On the other hand, if middle range data is not discarded when an arbitrary split is made for reasons other than ridding the data of nonlinearity, then pooling would appear to be a reasonable step toward restoring the information contained within the total bivariate set of data.

Implications of this study relate to the use of the pooled within-subgroup correlation coefficient and to generalizations of this study in terms of pooling multiple subgroups in the calculation of correlation

TABLE 3  
Values of  $r_1$  Resulting from One Pattern of Subgroup Centroid  
Difference and Five Total Group Sample Sizes:  $r_w = 0.0000$

Total group sample size	Significance levels for $t$ distribution when subgroup mean differences equal critical values			
	$p < .05$	$p < .01$	$p < .001$	$p < .0001$
Pattern 1: $\bar{X}_2 > \bar{X}_1$ , $\bar{Y}_2 > \bar{Y}_1$ , both significant				
10	.3993	.5846	.7606*	.8637**
50	.0777	.1303	.2038	.2728
100	.0386	.0658	.1051	.1438
200	.0193	.0330	.0533	.0738
1000	.0038	.0066	.0108	.0151

Note.—Asterisks refer to significance levels of unit-normal  $z$  tests comparing  $r_1$  and  $r_w$ .

\*  $p < .05$ .

\*\*  $p < .01$ .

matrices. Assuming no difference between the within-subgroup correlation coefficients (non-rejection of the first stage test), the pooled within-subgroup correlation coefficient is useful as a descriptive statistic with hypothesis-testing capabilities resulting from its equivalence to a first-order partial correlation coefficient. In research involving multiple dependent variables that are analyzed via several  $t$  tests comparing means, rather than a one-way two cell multiple analysis of variance, intercorrelations among the dependent variables should be assessed through pooled within-subgroup correlation coefficients rather than total group correlation coefficients. Otherwise, the meaningfulness of the intercorrelations would be contingent upon the failure to find cell differences in all of the  $t$  tests.

Considering the demonstrated varieties of possible distortion of a single correlation coefficient from the haphazard pooling of only two heterogeneous subgroups, the generalizations of these results must be inherently more complex, i.e., the effects of pooling multiple subgroups on correlation matrices. If, indeed, such complex distortion can be demonstrated, and it is desirable to pool data for reasons of parsimony, this suggests that further study should be devoted to the multivariate case and the development of appropriate test procedures.

## REFERENCES

- Dunlap, J. W. Combinative properties of correlation coefficients. *Journal of Experimental Education*, 1937, 5, 286-288.
- Games, P. A. and Klare, G. R. *Elementary statistics: Data analysis for the behavioral sciences*. New York: McGraw-Hill, 1967.
- Glass, G. V and Stanley, J. C. *Statistical methods in education and psychology*. Englewood Cliffs, N. J.: Prentice-Hall, 1970.
- Guilford, J. P. *Fundamental statistics in psychology and education*. (4th ed.) New York: McGraw-Hill, 1965.
- Lindquist, E. F. *Statistical analysis in educational research*. Boston: Houghton Mifflin, 1940.
- Pearson, K. Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. *Philosophical Transactions of the Royal Society (Series A)*, 1896, 187, 253-318.
- Pearson, K., Lee, A., and Bramley-Moore, L. Genetic (reproductive) selection: Inheritance of fertility in man and of fecundity in thoroughbred racehorses. *Philosophical Transactions of the Royal Society (Series A)*, 1899, 192, 257-330.
- Sockloff, A. L. and Edney, J. N. Some extensions of Student's  $t$  and Pearson's  $r$  central distributions. *Technical Report 72-5*. Philadelphia: Measurement and Research Center, Temple University, 1972.
- Walker, H. M. and Lev, J. *Elementary statistical methods*. (3rd ed.) New York: Holt, Rinehart, and Winston, 1969.
- Yule, G. U. and Kendall, M. G. *An introduction to the theory of statistics*. (11th ed.) London: Griffin, 1937.

## THE $r$ -POINT BISERIAL LIMITATION

R. A. KARABINUS

Northern Illinois University

A study was made of the  $r$ -point biserial coefficient using four non-normal distributions for the continuous variable: rectangular, bimodal-normal, bimodal-peaked, and bimodal-peaked and skewed.  $N$ s of 10, 30, and 100 were used. It was argued that linearity was the main assumption required when using Pearson correlations and that the usual maximum  $r$ -point biserial of .798 could be exceeded when the shape of the continuous variable more nearly approached that of the dichotomized variable. Correlations were found over .80 with rectangular distributions, and over .90 with bimodal-peaked distributions.

It is generally accepted by many practitioners of educational statistical techniques that the  $r$ -point biserial correlation coefficient has a limitation of about .80. This is the limitation that most educational statistics books cite, and it is based on the assumption that the  $Y$  variable (e.g., total test scores in an item analysis) is normally distributed and the  $X$  variable (e.g., scores on dichotomous items, 1 or 0) is a true dichotomy, split in such a way that  $p = q = .5$ . Since the  $r$ -point biserial is a member of the Pearson product moment correlation family, the other assumptions of linearity and homoscedasticity are also mentioned. However, a different interpretation of the normality assumption was taken by Walker and Lev (1953), who indicated that the assumption of normality is applicable for each set of  $Y$  scores paired with each of the  $X$  values. This would give a bimodal distribution on the  $Y$  variable. Since it is possible to obtain correlations above .80 with a bimodal distribution on the  $Y$  variable (with the  $X$  variable split so that  $p = q = .5$ ), the question of the traditionally accepted limitation and assumptions of the  $r$ -point biserial needs to be reopened.



Even though Nefzger and Drasgow (1957) made a strong case for not needing to meet the assumptions of normality when using any of the Pearson product moment correlations, most authors of beginning statistics books still cite the triumvirate: linearity, normality, and homoscedasticity. If, indeed, linearity is the chief assumption to be met with the Pearson coefficients, as suggested by Nefzger and Drasgow, then the other two conditions may accompany linearity but they are not necessary conditions. It is quite possible to have both linearity and homoscedasticity without having normal distributions, as long as the two variables have the same shape (both in skewness and kurtosis). This can be easily demonstrated by the interested reader by plotting a scattergram of two negatively (or positively) skewed variables that are moderately to highly related.

Applying these assumptions to the fourfold point correlation coefficient ( $\phi$ ), another member of the Pearson family, raises similar concerns. Both Glass and Stanley (1970) and McNemar (1962) explained the need to have similar distributions on the dichotomous variables in order to obtain the best estimates of correlations. Linearity was the main concern, and the authors showed that when linearity is violated by having disproportionate ratios of 1's and 0's for each variable, there is a tenuous effect on the correlation coefficient.

If normality is not a requirement for the  $\phi$  coefficient, then why should it be for the  $r$ -point biserial? To ensure linearity, however, it is necessary to have the same shape for each variable, both in skewness and kurtosis, especially the former. In light of this, the usually cited limitation of .80 for the  $r$ -point biserial can be understood. One variable is continuous and normal and the other dichotomous and symmetrical (if  $p = q$ ). When the dichotomous variable is split at the median, it is as "normal" as it can be, and the highest possible coefficient that can be obtained is .798.

If it is appropriate to use the Pearson coefficients when the variables are both linear and similar in shape, as suggested above, then if the shape of the continuous variable were made more similar to that of the dichotomous variable, the .798 limitation should not hold. Bowers (1972) found coefficients of .866 for rectangular distributions of the continuous variable with equal non-overlapping distributions for the dichotomous variable. He also found coefficients up to .849 when the continuous variable was skewed and the dichotomous variable split at .6 and .4. In an earlier study, Adams (1960) reported a  $r$ -point biserial of .839 with a platykurtic distribution using an  $n$  of 512 and an equal split on the dichotomous variable.

It is clear to the author that the main reason for the limiting value of the  $r$ -point biserial is the lack of similarity of distribution in the two

variables. While it is recognized that it is impossible to obtain a perfect correlation with the  $r$ -point biserial (this can occur only with two continuous variables or two dichotomous variables), it is possible to more nearly approach  $\pm 1.00$  by making the shape of the continuous variable more like the shape of the dichotomous one. This can be done by having a bimodal but symmetrical distribution on the continuous variable, which for each part of the dichotomous variable would be as peaked as possible.

In this study,  $r$ -point biserial coefficients were calculated using five different shapes of the continuous variable for comparative purposes:

Normal: distributions were less than perfectly normal because of the number of scale units used (see Table 2) and the size of  $n$ .

No fractional frequencies were used.

Rectangular: perfectly flat distributions. The coefficients were a function of the number of scale units used.

Bimodal-normal: the distribution of the continuous variable was made as normal as possible for *each* of the dichotomized values without using fractional frequencies.

Bimodal-peaked: for *each* dichotomized value, the distribution of the continuous variable remained symmetrical but was made as peaked (leptokurtic) as possible.

Bimodal-peaked and skewed: same as above (bimodal-peaked) except each of the distributions was highly negatively skewed.

Three different  $n$ 's were chosen, 10, 30, and 100 (with slight adjustment for the rectangular distribution), with two different  $p$  values, .5 and .6. One set of coefficients was calculated when there was no overlap of the continuous variable from one of the dichotomized values to the other, and another set with slight overlap (See Table 2 for the actual frequencies and overlap used). Table 1 gives the resulting  $r$ -point biserial coefficients under the described circumstances, and Table 2 gives the actual distributions used over each of the dichotomized values.

The calculated coefficients shown in Table 1 clearly indicate that  $r$ -point biserial values can be found above .798 when the shape of the continuous variable is more like that of the dichotomized one. (Those coefficients  $> .798$  under the "Normal distribution" occurred because the distributions were not perfectly normal.) In practical situations, researchers seldom will replicate the identical conditions assumed in this study, but they may approach them. For example, platykurtic and occasionally bimodal distributions do occur in item analyses of teacher-made tests. Therefore, it might be helpful to know that the  $r$ -point biserial, as one of the Pearson product mo-

TABLE 1  
*Point Biserial Coefficients with Different Shapes of Continuous Variable*

<i>n</i>	<i>Y</i> on <i>X</i>	Normal <i>p</i> = .5	Normal <i>p</i> = .6	Rectangular <i>p</i> = .5	Rectangular <i>p</i> = .6	Bimodal-normal <i>p</i> = .5	Bimodal-normal <i>p</i> = .6	Bimodal-peaked <i>p</i> = .5	Bimodal-peaked <i>p</i> = .6	Bimodal-peaked skewed <i>p</i> = .5	Bimodal-peaked skewed <i>p</i> = .6
10	No Overlap	.809	.539	.878	.866	.921	.919	.921	.919	.882	.896
	Overlap	.730	.745	.849	.722	.845	.840	.845	.840	.781	.813
30	No Overlap	.805	.817	.870	.853	.927	.919	.951	.934	.936	.916
	Overlap	.696	.772	.857	.833	.889	.889	.926	.909	.904	.871
100	No Overlap	.802	.706	.870	.853	.950	.948	.978	.977	.960	.960
	Overlap	.796	.769	.861	.803	.934	.934	.970	.968	.946	.946

TABLE 2

[illegible]

ment correlation coefficients, has no limitation that is not present in any of the other members of the Pearson family (except for the natural limitation of the variables themselves). While it is not possible to obtain the perfect  $\pm 1.00$ , one can approach it, e.g., .978 with an  $n$  of 100, with a bimodal-peaked distribution on the continuous variable and no overlap on the dichotomized variable.

In summary, with  $n$ 's  $\geq 30$  and rectangular distributions of the continuous variable having little or no overlap on each of the dichotomized values, coefficients above .80 can be expected. Similarly, with bimodal distributions having little or no overlap on each of the dichotomized values, coefficients of .90 or above can be expected.

## REFERENCES

- Adams, J. F. The effect of non-normally distributed criterion scores on item analysis techniques. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1960, 20, 317-320.
- Bowers, J. A note on comparing  $r$ -biserial and  $r$ -point biserial. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1972, 32, 771-775.
- Glass, G. V and Stanley, J. C. *Statistical methods in education and psychology*. Englewood Cliffs, N. J.: Prentice-Hall, 1970.
- McNemar, Q. *Psychological statistics* (3rd ed.). N.Y.: John Wiley and Sons, 1962.
- Nefzger, M. D. and Drasgow, J. The needless assumption of normality in Pearson's  $r$ . *American Psychologist*, 1957, 12, 623-625.
- Walker, H. M. and Lev, J. *Statistical inference*. N. Y.: Holt, Rinehart and Winston, 1953.



## ON SPLITTING THE TAILS UNEQUALLY: A NEW PERSPECTIVE ON ONE-VERSUS TWO-TAILED TESTS

SANFORD L. BRAVER<sup>1</sup>  
Arizona State University

The controversy that has raged since the early fifties regarding the admissibility of one-tailed tests of hypotheses was examined. From the review of that literature, it was concluded that the main advantage of the one-tailed test was the gain in power for the prediction while its main disadvantage was its inability to test for significance if the results were opposite to prediction. It is argued here that splitting  $\alpha$  unequally between the two tails, placing most of the rejection region on the side of the prediction but a smaller fraction on the opposite side provides both power and the ability to detect opposite-to-prediction outcomes. This compromise procedure requires a finer choice in the splitting of  $\alpha$  than the dichotomous choice of putting either all or exactly half of  $\alpha$  in the favored tail, i.e., the choice between a one- or a two-tailed test. Rules for the most effective split, based on Bayesian considerations, are prescribed. The fraction of  $\alpha$  in the predicted tail should be equal to the investigator's a priori probability that the predicted order, as opposed to the reversed order, of sample means will be obtained. A table of  $t$ -values is presented which gives critical regions for significance, both "expected" and "unexpected," at specified levels of a priori probability.

SINCE the early fifties lengthy debates regarding the propriety of one-versus two-tailed tests of hypotheses have appeared in the literature. The sole area of agreement appears to be that the two-tailed test is appropriate when the investigator is concerned merely with the presence or absence of the effect of some two-level independent variable on the

---

<sup>1</sup> Requests for reprints should be sent to: Dr. Sanford L. Braver, Department of Psychology, Arizona State University, Tempe, Arizona, 85281.

dependent variable,<sup>2</sup> without being concerned, or hypothesizing in advance, as to the direction of difference. In this test, the null is that the two population means are identical ( $H_0: \mu_1 - \mu_2 = 0$ ), or more precisely, that the two samples are drawn from a single population. An improbably large difference in the two sample means, regardless of which is the greater, is taken as evidence opposed to this null and in favor of its alternative,  $H_a: \mu_1 - \mu_2 \neq 0$ . In order to assure that the probability of a Type I error (rejecting a true null hypothesis) is held at some fixed level,  $\alpha$ , we divide  $\alpha$  in half. Thus,  $.5\alpha$  proportion of the area on either extreme (or tail) of the sampling distribution of mean differences is taken as the "rejection of null" region.

Many authorities (Marks, 1951, 1953; Jones, 1952, 1954) argued that a different kind of logic is applicable when the investigator hypothesizes in advance which of the two means should be the larger, by far the most usual occurrence in psychological research. Here, the null is termed directional, and takes the form  $H_0: \mu_1 - \mu_2 \leq 0$ , while the alternative becomes  $H_a: \mu_1 - \mu_2 > 0$ . Again only evidence in favor of the alternative hypothesis, that is finding that the obtained order of means matches the order predicted by  $H_a$ , can be used to reject the null. Thus, the entire  $\alpha$ , rather than half of it, is placed on the side of the null-generated sampling distribution corresponding to the ordering of means predicted by  $H_a$ , and becomes the rejection region. This one-tailed test procedure has the effect of requiring less extreme mean differences in the direction predicted than does the two-tailed test to reject the null with probability  $\alpha$  of a Type I error. The latter is referred to as a gain in power.

The advice to use one-tailed tests is not without severe critics, however. In particular, Hick (1952) and Burke (1953, 1954) have joined the fray opposing this use. Their most compelling point is that this procedure prevents the investigator from evaluating for significance differences in the *unexpected* direction, no matter how large. Inasmuch as the insights to be gained from clear-cut results which are counter-intuitive or counter-theoretical sometimes exceed in importance those from findings that are consistent with predictions, this short-coming is crucial.

---

<sup>2</sup> The argument being developed is cast in terms of mean differences or  $t$ -tests, but, of course, is equally applicable to any potentially two-directional test. Thus the argument for tests of central tendency using non-parametric indices (e.g., Mann-Whitney U, median split, etc.) tests of correlation coefficients different from zero, tests of differences in correlation coefficients, tests of chi-square with two levels on the predictor variables, and even  $F$ -tests in analysis of variance would all be exactly analogous. In the latter case, in a factorial (or one-way) design, any 1  $df$  test or contrast may be converted to one sensitive to direction by square-rooting the obtained  $F$ , which is then equivalent to a  $t$ -statistic with  $df$  equal to the  $df$  of the  $F$  denominator.

Kimmel (1957) took a slightly more moderate position. He wished to "limit the use of one-tailed tests to . . . infrequent situations" (p. 353) rather than to outlaw them entirely. In developing his criteria for their authorized use, however, he is clearly in agreement with Hick and Burke that most unexpected findings must be testable for significance. Thus these criteria essentially require the investigator to perform a two-tailed test except when differences in the opposite direction are impossible or meaningless.

Hence, the main disadvantage of the one-tailed test, as compared to the two-tailed, its inability to deal responsibly with results in the other direction from that hypothesized, is seen by its critics as fatal. However, the main advantage of the one-tailed test, additional power, is seen by its proponents as overriding this disadvantage. Fisher, in defining power as the probability of detecting true population differences, used the term synonymously with "precision." Furthermore, Overall (1969) proved that increasing power has the additional advantage of decreasing the proportion of significant results which are due to Type I error. Despite its obvious importance, power is a consideration which has been too often either neglected or misunderstood. Cohen (1962) has shown that the level of power attained by studies in the social-personality area is unacceptably low. And even sophisticated researchers have misguided faith in the ability of well-planned studies to surmount deficiencies of power, as Tversky and Kahneman (1971) have demonstrated.

The present argument accepts the rationality inherent in both positions and urges a compromise. Rather than taking a stance with regard to whether the one- or the two-tailed test is the most seriously flawed, a procedure is developed which can capitalize on the advantages of each.

The procedure that can mediate this compromise is the unequal splitting of  $\alpha$  between the two tails; for example,  $.8\alpha$  can be placed in the tail which represents the predicted direction, and  $.2\alpha$  in the other. By splitting  $\alpha$  unequally, the investigator is not forced to make a binary choice between the traditional two- or traditional one-tailed tests. Indeed, it is difficult to formulate a rationale for why these should be the only options available. By splitting  $\alpha$  unequally the investigator is instead free to choose from a continuous range of possibilities that split of  $\alpha$  which is optimum for his situation. This range extends from (and includes)  $.5\alpha$  in each tail, to all of  $\alpha$  in the favored tail, thus including the traditional tests as the boundary cases. In making this more flexible choice, the investigator realizes simultaneously the advantages of both the traditional tests. The researcher gains power on the side where he wants it, but he also retains the ability to detect unexpected results. The latter is purchased at the cost of some of

the former, but the investigator decides in advance the best balance for his particular research.

By dividing  $\alpha$  a priori into two unequal parts and assigning the greater fraction to the expected tail, the investigator avoids a difficult and embarrassing dilemma that frequently arises. This dilemma occurs when he is faced with unexpected results which would have been significant by a traditional two-tailed test, after having decided upon a one-tailed test. Goldfried (1959) noted that the investigator in this situation has available to him three options: (1) He can ignore the result by considering it part of the null hypothesis of no difference; (2) he can test the significance of the unexpected result with the same data, but this kind of "cheating" has the effect of inflating  $\alpha$ ; or (3) he can gather new data, testing the reliability of the unexpected finding with a reversed one-tailed test. All of these options have some obvious disadvantages. The proposed procedure minimizes or avoids these disadvantages. The investigator states in advance which region of the unexpected tail will constitute results he will consider synonymous with the conclusion of "no difference except that due to sampling fluctuation," and which will fit with the conclusion of "results significantly opposite to expected." Secondly, the two rejection regions, since they sum to  $\alpha$ , keep  $\alpha$  constant at its preordained level. Finally, both hypotheses—the expected one and the unexpected one—can be tested with the same data.<sup>3</sup>

### *Deciding on Fractions*

If the unequal splitting of  $\alpha$  is to be a viable option for the researcher, it is necessary to develop a rational basis for deciding upon the split of  $\alpha$  into fractions.

The first consideration is that the choice should obviously be made in advance of seeing the data. Though Hick (1952) argued that "it makes no difference when a theory or hypothesis is conceived . . . Logic is timeless," (p. 316), Marks (1953) correctly responds to this by noting that *how* it is conceived—from inspection of the set of data being analyzed or from theoretical considerations—is, however, of statistical importance.

<sup>3</sup> A clever approach to directional inference, suggested by Shaffer (1972) is to test simultaneously two sets of nulls and alternatives:  $1H_0: \mu_1 \geq \mu_2$  vs.  $1H_a: \mu_1 < \mu_2$ , and  $2H_0: \mu_1 \leq \mu_2$  vs.  $2H_a: \mu_1 > \mu_2$ . If the investigator tests each null at  $.5\alpha$ , the maximum probability of at least one false rejection is  $\alpha$ . This approach, too, is amenable to the suggestions made here. If the experimental prediction favored by the investigator is  $1H_a$ , he should inflate the  $\alpha$  for the test of  $1H_0$ , and correspondingly decrease it for the test of  $2H_0$ , such that the sum of the two  $\alpha$ 's equal the predetermined overall  $\alpha$ .



The second issue concerns the belief structure or conclusions drawn by the investigator, which, of course, materially affect his subsequent research activities and thus determine the quantity and nature of evidence that will be obtained in the future. With the present two-tailed test, a statistically significant finding that is directionally opposite to that which was theoretically predicted, does not appear to have the same implications—in terms of these conclusions or future research activities—as “expected significance.” It is true that opposite-to-predicted significant results are likely to weaken the investigator’s belief in his hypothesis while expected significant results will strengthen this belief. However, it is also true that the former is unlikely to strengthen his belief in the opposite hypothesis *as much as* the latter will strengthen his belief in the proposed hypothesis. This can be demonstrated by observing that investigators more often attempt replication of a study when the results are significant in the opposite direction than when they are significant in the direction predicted. What is empirically the case corresponds to what statistical decision (i.e., Bayesian) theory prescribes as optimum for the decision-making researcher.

The recommended flexible division of  $\alpha$  would enable us to deal with the asymmetry of the impact of expected vs. unexpected significance. The split of  $\alpha$  *should* be arranged in such a way that the investigator will be as convinced by “unexpected significance” as by “expected significance.” That is, the split must be arranged so that if a test statistic just falls in the smaller of the rejection regions the investigator’s belief in the validity of the reversed hypothesis should equal his belief in the predicted hypothesis if the test statistic had just fallen into the larger of the rejection regions. To clarify, and following Kaiser (1960), we divide the usual two hypotheses, null and alternative, into three which are again mutually exclusive and exhaustive:  $H_1: \mu_1 < \mu_2$ ;  $H_2: \mu_1 = \mu_2$ ; and  $H_3: \mu_1 > \mu_2$ . Suppose that  $H_3$  is the hypothesis favored by the theory, and the investigator had decided upon the classical two-tailed test, i.e., each tail contains  $.5\alpha$ . It is unlikely that a  $t$ -value which just falls into the  $H_1$  rejection area will confirm the investigator’s belief in the validity of  $H_1$  to the extent that a  $t$ -value which barely crosses the  $H_3$  rejection region does so for  $H_3$ . He is much more likely to attempt to explain the unexpected,  $H_1$ -consistent, results than the expected,  $H_3$ -consistent, results as chance deviation due to sampling (i.e., Type I) or measurement error. In contrast, he will typically accept the latter as confirmation. If one simply will not be convinced by an opposite-to-predicted result without further replication, he is in essence throwing away up to half of his  $\alpha$  if he places it in that tail. In terms of the present argument for keeping the



implications of achieving significance consistent for the researcher's conclusions, the 50-50 split is unwise. However, some other split of  $\alpha$  can be found where the investigator's belief changes *identically* for results in either rejection region. The task is to adjust the size of these two rejection regions in such a way as to *equalize* the investigator's temptation to regard results in either of these regions as arising from Type I error. This can be done by entertaining as possibilities various splits and with each asking the question: "If the  $t$ -value falls into the  $H_1$  rejection area will my conviction in the validity of  $H_1$  be equal to my conviction in the validity of  $H_3$  if the value falls into its rejection region?" If so, the correct split has been made; if not continual readjustments of the split are made until the answer is affirmative.

While the above procedure is tedious, another can be found which will have the same desired affect, i.e., of producing *identical* changes in belief for results in either rejection region. This more convenient procedure makes use of a version of Bayes' formula to the effect that proportionally greater evidence should be required to convince the rational decision-maker of the validity of a proposition he considers improbable than one which has high a priori probability. Thus if the a priori probability of  $H_3$  is four times that of  $H_1$ , the evidence for  $H_1$  in order to accept  $H_1$  should be required to be four times as compelling as the evidence for  $H_3$  in order to accept  $H_3$ . Thus the investigator should compare  $H_3$  to  $H_1$  for a priori probability.<sup>4</sup> The fraction of  $\alpha$  placed in the larger tail should equal the a priori probability of  $H_3$ .

Yet another mathematical relationship can be utilized to make the procedure still more concrete. The investigator's a priori probability of  $H_3$  should be equivalent to his judgment of the likelihood that the *sample means* will be in the order predicted by the theory rather than the reverse direction, ignoring for the moment the issue of significance. If, for example, the researcher believes, on the basis of previous accumulated evidence and all else that he knows of the situation, that  $\bar{X}_1$  is four times as likely to exceed  $\bar{X}_2$  as the reverse, i.e., his personal probability that  $\bar{X}_1 > \bar{X}_2$  equals .8, he should split  $\alpha$  so that  $.8\alpha$  lies in the  $H_3$  tail,  $.2\alpha$  in the  $H_1$  tail.

To ascertain whether .8 actually is his a priori probability he should ask himself whether he would be willing to accept *either* side of a money bet that  $\bar{X}_1$  will exceed  $\bar{X}_2$ , as opposed to  $\bar{X}_1 < \bar{X}_2$ , at 4 to 1 odds.<sup>5</sup> The basic rule becomes: *the proportion of  $\alpha$  allocated to the favored*

<sup>4</sup> This assumes that the a priori probability of  $H_2$  (that  $\mu_1$  equals exactly  $\mu_2$ ) is zero, a realistic assumption. See Bakan (1966), Meehl (1967).

<sup>5</sup> To offer a formula, and an additional example if the investigator would give A to B odds, where A is the larger, since his a priori probability is  $A/(A + B)$ , he should put  $(A/(A + B))\alpha$  in the larger tail,  $(B/(A + B))\alpha$  in the smaller. With odds of 3 to 2,  $A = 3$ ,  $B = 2$  and probability = .6.

*tail should equal the investigator's a priori probability that the sample means will be ordered as predicted*, for then he will be equally tempted to regard results in either rejection region as arising from Type I error.

Traditional one- and two-tailed tests emerge as special cases of this rule. A one-tailed test is performed when the a priori probability of reversal of means is 0, that is, when it is seen as impossible. (cf. Kimmel, 1957). A two-tailed test (50-50 split) is performed when the researcher has no a priori evidence which leads him to expect that  $\bar{X}_1$  is more likely to exceed  $\bar{X}_2$  than the reverse.

Table 1 presents critical values for the  $t$ -statistic, with  $\alpha = .05$ , in which  $\alpha$  is split in accordance with the a priori probability of the predicted direction of sample means. Two values are given for each combination of  $df$  with a priori probability. The "+" value corresponds to the  $t$ -value needed to reject the  $H_2$  null in favor of the predicted direction, while the "-" value presents the value necessary to reject the  $H_2$  null in favor of the reversed direction hypothesis. The conditional probability, given  $H_2$ , that  $t$  will exceed the "+" value or be less than the "-" value, equals .05. When the a priori probability is 1.0, the "+" values correspond to those needed for traditional one-tailed tests, while the "-" values are infinite. When the a priori probability is .5, the values for both "+" and "-" outcomes are identical to two-tailed test values. As another example using the case given earlier, where the odds were 4 to 1, the a priori probability would equal .8. Thus the "+" value corresponds to  $.8\alpha$  or .04, while the "-" value is for  $.2\alpha$  or .01. Given  $df$  of, for example, 21, if the sample means are ordered as predicted, an obtained value for  $t$  would have to equal or exceed 1.840 to claim "expected significance." The critical value for "unexpected significance," for the case when the order is reversed, is 2.518. By adopting these critical values, expected and unexpected significance should have equal implications for the researcher's a posteriori beliefs.

### *Discussion*

Splitting the tails unequally on the basis of the a priori probability that the predicted direction will obtain for the sample means calls into focus a point which remained obscured utilizing the traditional procedures. When a prediction of an experimental outcome is made, it follows from a combination of intuitive notions, informal observations, empirically untested theory, empirically tested theory, previous tangential research with varying outcomes, previous research with the present paradigm with varying outcomes, etc. Another consideration is the skill of the investigator in operationalizing his concepts and provid-

TABLE I  
Critical *t* Levels  
*A Priori Probability of Predicted Direction*

<i>df</i>	+/-	1.00	0.99	0.98	0.97	0.96	0.95	0.90	0.85	0.80	0.75
2	+	2.920	2.937	2.954	2.972	2.990	3.009	3.104	3.207	3.320	3.443
2	-	∞	-31.599	-22.327	-18.216	-15.764	-14.089	-9.925	8.073	-6.965	-6.205
3	+	2.353	2.364	2.376	2.387	2.398	2.410	2.471	2.536	2.605	2.681
3	-	∞	-12.924	-10.215	-8.891	-8.053	-7.453	-5.841	-5.047	-4.541	-4.177
4	+	2.132	2.141	2.150	2.159	2.168	2.178	2.226	2.278	2.333	2.392
4	-	∞	-8.610	-7.173	-6.435	-5.951	-5.597	-4.604	-4.088	-3.747	-3.495
5	+	2.015	2.023	2.031	2.039	2.047	2.055	2.098	2.143	2.191	2.242
5	-	∞	-6.869	-5.893	-5.376	-5.030	-4.773	-4.032	-3.634	-3.365	-3.163
6	+	1.943	1.950	1.958	1.965	1.973	1.980	2.019	2.060	2.104	2.151
6	-	∞	-5.959	-5.207	-4.800	-4.525	-4.317	-3.707	-3.372	-3.143	-2.969
7	+	1.895	1.901	1.908	1.915	1.922	1.929	1.966	2.005	2.046	2.090
7	-	∞	-5.408	-4.785	-4.442	-4.207	-4.029	-3.499	-3.203	-2.998	-2.841
8	+	1.860	1.866	1.872	1.879	1.886	1.893	1.928	1.965	2.004	2.046
8	-	∞	-5.041	-4.501	-4.199	-3.991	3.832	3.355	-3.085	-2.896	-2.752
9	+	1.833	1.839	1.846	1.852	1.858	1.865	1.899	1.935	1.972	2.013
9	-	∞	-4.781	-4.297	-4.024	-3.835	-3.690	-3.250	-2.998	-2.821	-2.685

10	+	1.812	1.819	1.825	1.831	1.837	1.844	1.877	1.911	1.948	1.987
10	-	$\infty$	-4.587	-4.144	-3.892	-3.716	-3.581	-3.169	-2.932	-2.764	-2.634
11	+	1.796	1.802	1.808	1.814	1.820	1.827	1.859	1.892	1.928	1.966
11	-	$\infty$	-4.437	-4.025	-3.789	-3.624	-3.497	-3.106	-2.879	-2.718	-2.593
12	+	1.782	1.788	1.794	1.800	1.806	1.812	1.844	1.877	1.912	1.949
12	-	$\infty$	-4.318	-3.930	-3.706	-3.550	-3.428	-3.055	-2.836	-2.681	-2.560
13	+	1.771	1.777	1.782	1.788	1.795	1.800	1.832	1.864	1.899	1.935
13	-	$\infty$	-4.221	-3.852	-3.639	-3.489	-3.372	-3.012	-2.801	-2.650	-2.533
14	+	1.761	1.767	1.773	1.778	1.785	1.790	1.821	1.854	1.887	1.923
14	-	$\infty$	-4.140	-3.787	-3.582	-3.438	-3.326	-2.977	-2.771	-2.624	-2.510
15	+	1.753	1.759	1.764	1.770	1.776	1.782	1.812	1.844	1.878	1.913
15	-	$\infty$	-4.073	-3.733	-3.535	-3.395	-3.286	-2.947	-2.746	-2.602	-2.490
16	+	1.746	1.752	1.757	1.763	1.769	1.775	1.805	1.836	1.869	1.904
16	-	$\infty$	-4.015	-3.686	-3.494	-3.358	-3.252	-2.921	-2.724	-2.583	-2.473
17	+	1.740	1.745	1.751	1.756	1.762	1.768	1.798	1.829	1.862	1.897
17	-	$\infty$	-3.965	-3.646	-3.459	-3.326	-3.222	-2.898	-2.706	-2.567	-2.458
18	+	1.734	1.740	1.745	1.751	1.757	1.762	1.792	1.823	1.855	1.890
18	-	$\infty$	-3.922	-3.610	-3.428	-3.298	-3.196	-2.878	-2.689	-2.552	-2.445

TABLE 1 (Continued)

<i>df</i>	+/—	0.70	0.67	0.65	0.60	0.58	0.56	0.54	0.52	0.50
19 +	1.729	1.735	1.740	1.746	1.751	1.757	1.786	1.817	1.850	1.884
19 -	∞	-3.883	-3.579	-3.401	-3.273	-3.174	-2.861	-2.674	-2.539	-2.433
20 +	1.725	1.730	1.736	1.741	1.747	1.753	1.782	1.812	1.844	1.878
20 -	∞	-3.850	-3.552	-3.376	-3.251	-3.153	-2.845	-2.661	-2.528	-2.423
21 +	1.721	1.726	1.732	1.737	1.743	1.748	1.777	1.808	1.840	1.873
21 -	∞	-3.819	-3.527	-3.355	-3.231	-3.135	-2.831	-2.649	-2.518	-2.414
22 +	1.717	1.723	1.728	1.734	1.739	1.745	1.773	1.803	1.835	1.869
22 -	∞	-3.792	-3.505	-3.335	-3.214	-3.119	-2.819	-2.639	-2.508	-2.405
23 +	1.714	1.719	1.725	1.730	1.736	1.741	1.770	1.800	1.832	1.865
23 -	∞	-3.768	-3.485	-3.318	-3.198	-3.104	-2.807	-2.629	-2.500	-2.398
24 +	1.711	1.716	1.722	1.727	1.733	1.738	1.767	1.797	1.828	1.861
24 -	∞	-3.745	-3.467	-3.301	-3.183	-3.091	-2.797	-2.620	-2.492	-2.391
25 +	1.708	1.713	1.719	1.724	1.730	1.735	1.764	1.793	1.825	1.858
25 -	∞	-3.725	-3.450	-3.287	-3.170	-3.078	-2.787	-2.612	-2.485	-2.385
26 +	1.706	1.711	1.716	1.722	1.727	1.733	1.761	1.791	1.822	1.855
26 -	∞	-3.707	-3.435	-3.274	-3.158	-3.067	-2.779	-2.605	-2.479	-2.379
27 +	1.703	1.709	1.714	1.719	1.725	1.730	1.758	1.788	1.819	1.852
27 -	∞	-3.690	-3.421	-3.261	-3.147	-3.057	-2.771	-2.598	-2.473	-2.373



28	+	1.701	1.706	1.712	1.717	1.723	1.728	1.756	1.786	1.817	1.849
28	-	$\infty$	-3.674	-3.408	-3.250	-3.136	-3.047	-2.763	-2.592	-2.467	2.368
29	+	1.699	1.704	1.710	1.715	1.720	1.726	1.754	1.783	1.814	1.847
29	-	$\infty$	-3.659	-3.396	-3.239	-3.127	-3.038	-2.756	-2.586	-2.462	-2.364
30	+	1.697	1.703	1.708	1.713	1.719	1.724	1.752	1.781	1.812	1.844
30	-	$\infty$	-3.646	-3.385	-3.230	-3.118	-3.030	-2.750	-2.580	-2.457	-2.360
40	+	1.684	1.689	1.694	1.699	1.705	1.710	1.737	1.766	1.796	1.828
40	-	$\infty$	-3.551	-3.307	-3.160	-3.055	-2.971	-2.704	-2.542	-2.423	-2.329
60	+	1.671	1.676	1.681	1.686	1.691	1.696	1.723	1.751	1.781	1.812
60	-	$\infty$	-3.460	-3.232	-3.094	-2.994	-2.915	-2.660	-2.504	-2.390	-2.299
120	+	1.658	1.663	1.668	1.673	1.678	1.683	1.709	1.737	1.766	1.796
120	-	$\infty$	-3.373	-3.160	-3.030	-2.935	-2.860	-2.617	-2.468	-2.358	-2.270
$\infty$	+	1.645	1.649	1.654	1.659	1.664	1.669	1.695	1.723	1.751	1.781
$\infty$	-	$\infty$	-3.291	-3.090	-2.967	-2.878	-2.807	-2.576	-2.433	-2.326	-2.241
2	+		3.578	3.666	3.728	3.896	3.969	4.046	4.127	4.212	4.303
2	-		-5.643	-5.368	-5.204	-4.849	-4.724	-4.609	-4.500	-4.398	-4.303
3	+		2.763	2.815	2.852	2.951	2.993	3.037	3.083	3.132	3.182
3	-		-3.896	-3.755	-3.670	-3.482	-3.415	-3.352	-3.292	-3.236	-3.182

TABLE 1 (Continued)

df	+/-	0.70	0.67	0.65	0.60	0.58	0.56	0.54	0.52	0.50
4	+	2.456	2.497	2.525	2.601	2.633	2.667	2.702	2.738	2.776
4	-	-3.298	-3.197	-3.135	-2.999	-2.949	-2.903	-2.858	-2.816	-2.776
5	+	2.297	2.333	2.357	2.422	2.449	2.477	2.507	2.538	2.571
5	-	-3.003	-2.920	-2.870	-2.757	-2.715	-2.677	-2.640	-2.604	-2.571
6	+	2.201	2.233	2.255	2.313	2.338	2.364	2.390	2.418	2.447
6	-	-2.829	-2.757	-2.712	-2.612	-2.576	-2.541	-2.508	-2.477	-2.447
7	+	2.136	2.166	2.187	2.241	2.264	2.288	2.312	2.338	2.365
7	-	-2.715	-2.649	-2.608	-2.517	-2.483	-2.452	-2.421	-2.392	-2.365
8	+	2.090	2.118	2.138	2.189	2.211	2.233	2.257	2.281	2.306
8	-	-2.634	-2.572	-2.535	-2.449	-2.417	-2.388	-2.359	-2.332	-2.306
9	+	2.055	2.082	2.101	2.150	2.171	2.193	2.215	2.238	2.262
9	-	-2.574	-2.516	-2.480	-2.398	-2.369	-2.340	-2.313	-2.287	-2.262
10	+	2.028	2.055	2.072	2.120	2.140	2.161	2.183	2.205	2.228
10	-	-2.527	-2.472	-2.437	-2.359	-2.331	-2.303	-2.277	-2.252	-2.228
11	+	2.006	2.032	2.050	2.096	2.116	2.136	2.157	2.179	2.201
11	-	-2.491	-2.437	-2.404	-2.328	-2.300	-2.274	-2.249	-2.224	-2.201
12	+	1.989	2.014	2.031	2.076	2.096	2.115	2.135	2.157	2.179
12	-	-2.461	-2.408	-2.376	-2.303	-2.276	-2.250	-2.225	-2.202	-2.179

13	+	1.974	1.999	2.016	2.060	2.079	2.098	2.118	2.139	2.160
13	-	-2.436	-2.385	-2.353	-2.282	-2.255	-2.230	-2.206	-2.183	-2.160
14	+	1.962	1.986	2.002	2.046	2.065	2.084	2.103	2.124	2.145
14	-	-2.415	-2.365	-2.334	-2.264	-2.238	-2.213	-2.189	-2.167	-2.145
15	+	1.951	1.975	1.991	2.034	2.052	2.071	2.091	2.111	2.131
15	-	-2.397	-2.348	-2.318	-2.249	-2.223	-2.199	-2.175	-2.153	-2.131
16	+	1.942	1.965	1.981	2.024	2.042	2.060	2.080	2.099	2.120
16	-	-2.382	-2.333	-2.304	-2.235	-2.210	-2.186	-2.163	-2.141	-2.120
17	+	1.934	1.957	1.973	2.015	2.033	2.051	2.070	2.090	2.110
17	-	-2.368	-2.321	-2.291	-2.224	-2.199	-2.175	-2.153	-2.131	-2.110
18	+	1.926	1.949	1.965	2.007	2.025	2.043	2.062	2.081	2.101
18	-	-2.356	-2.309	-2.280	-2.214	-2.189	-2.166	-2.143	-2.122	-2.101
19	+	1.920	1.943	1.959	2.000	2.017	2.035	2.054	2.073	2.093
19	-	-2.346	-2.299	-2.271	-2.205	-2.180	-2.157	-2.135	-2.113	-2.093
20	+	1.914	1.937	1.953	1.994	2.011	2.029	2.047	2.066	2.086
20	-	-2.336	-2.290	-2.262	-2.197	-2.173	-2.150	-2.128	-2.106	-2.086
21	+	1.909	1.932	1.947	1.988	2.005	2.023	2.041	2.060	2.080
21	-	-2.328	-2.282	-2.254	-2.189	-2.166	-2.143	-2.121	-2.100	-2.080

TABLE 1 (Continued)

<i>df</i>	+/ -	0.70	0.67	0.65	0.60	0.58	0.56	0.54	0.52	0.50
22	+	1.905	1.927	1.942	1.983	2.000	2.018	2.036	2.054	2.074
22	-	-2.320	-2.275	-2.247	-2.183	-2.159	-2.137	-2.115	-2.094	-2.074
23	+	1.900	1.922	1.938	1.978	1.995	2.013	2.031	2.049	2.069
23	-	-2.313	-2.268	-2.240	-2.177	-2.153	-2.131	-2.109	-2.089	-2.069
24	+	1.896	1.919	1.934	1.974	1.991	2.008	2.026	2.045	2.064
24	-	-2.307	-2.262	-2.235	-2.172	-2.148	-2.126	-2.104	-2.084	-2.064
25	+	1.893	1.915	1.930	1.970	1.987	2.004	2.022	2.040	2.060
25	-	-2.301	-2.257	-2.229	-2.167	-2.143	-2.121	-2.100	-2.079	-2.060
26	+	1.890	1.912	1.927	1.967	1.983	2.000	2.018	2.037	2.056
26	-	-2.296	-2.252	-2.225	-2.162	-2.139	-2.117	-2.095	-2.075	-2.056
27	+	1.887	1.909	1.924	1.963	1.980	1.997	2.015	2.033	2.052
27	-	-2.291	-2.247	-2.220	-2.158	-2.135	-2.113	-2.092	-2.071	-2.052
28	+	1.884	1.906	1.921	1.960	1.977	1.994	2.011	2.030	2.048
28	-	-2.286	-2.243	-2.216	-2.154	-2.131	-2.109	-2.088	-2.068	-2.048
29	+	1.881	1.903	1.918	1.957	1.974	1.991	2.008	2.026	2.045
29	-	-2.282	-2.239	-2.212	-2.150	-2.128	-2.106	-2.085	-2.065	-2.045
30	+	1.879	1.901	1.915	1.955	1.971	1.988	2.006	2.024	2.042
30	-	-2.278	-2.235	-2.208	-2.147	-2.124	-2.103	-2.082	-2.062	-2.042

40	+	1.862	1.883	1.897	1.936	1.952	1.968	1.985	2.003	2.021
40	-	-2.250	-2.209	-2.183	-2.123	-2.101	-2.080	-2.059	-2.040	-2.021
60	+	1.845	1.865	1.880	1.917	1.933	1.949	1.965	1.983	2.000
60	-	-2.223	-2.182	-2.157	-2.099	-2.078	-2.057	-2.038	-2.019	-2.000
120	+	1.828	1.848	1.862	1.899	1.914	1.930	1.946	1.963	1.980
120	-	-2.196	-2.157	-2.132	-2.076	-2.055	-2.036	-2.016	-1.998	-1.980
$\infty$	+	1.812	1.832	1.845	1.881	1.896	1.911	1.927	1.943	1.960
$\infty$	-	-2.170	-2.135	-2.109	-2.054	-2.034	-2.015	-1.996	-1.978	-1.960



ing a fair test. However, no two experimental predictions combine these ingredients in precisely the same way. Some predictions are made with extreme confidence and vigor, and others with great timidity. To permit a choice only between a one- and a two-tailed test is to disallow finer distinctions in the strength or firmness of predictions made in different studies. The continuous choice of splits is superior, not only because of salutary effects on the conduct of research, but also because it allows the researcher to summarize in a single index, i.e., the a priori probability statement, all the considerations which led him to his prediction. While an objective formula for combining the considerations into a probability statement would, of course, be preferable to the subjective operation advocated here, it is unfortunately unrealistic.

It should be emphasized, however, that the personal probability statement should be a public one which the researcher can defend by reference to previous literature, etc. Indeed, the typical introductory section of most research reports is a listing of the considerations which led the research to his prediction, and could readily and logically culminate in the statement of the a priori probability used in the test of the hypothesis. Readers (or editors) whose personal probability differed from the investigator's would be thus free to draw different implications from the evidence.

One apparent drawback to the proposed procedure is its vulnerability to abuse. Suppose an investigator, in advance of seeing the data, decided upon a 60-40 split, i.e., his subjective probability was .6 that  $\bar{X}_1$  would exceed  $\bar{X}_2$ . Assume, with  $df = 16$ , his obtained  $t$ -value was +1.90. From the table he observes that he has not obtained significance in the expected direction. If he had instead chosen .8 as his a priori probability it is clear that significance *would have* been achieved. What is to prevent him from being seduced into discarding his earlier split and deciding that confidence of .8 had indeed been warranted? On reflection it can be recognized that the same argument applies with equal force to switching from a two- to a one-tailed test to achieve significance. As in the latter we must rely on the integrity of the investigator.

There is no doubt that the compromise procedure of splitting the tails unequally will offend some traditionalists, especially hardcore advocates of two-tailed tests. Clearly this compromise is preferable to the overuse of the one-tailed test. Nonetheless, a basic objection might be that such a procedure formalizes and thus sanctions the custom of giving less credence to disconfirming than confirming evidence, a practice which could be viewed as scientifically corrupt. It is perhaps a sufficient defense of this compromise procedure to argue that the above characterizes the way all-too-human investigators presently act

anyhow, however much we might prefer some more ideal behavior. Science progresses by what scientists do, not by what we would like them to do. The procedure here advocated provides a set of operations, and the rationale, to do what is already done, but in a far more systematic and orderly fashion.

A stronger and less cynical defense of the procedure, however, can be offered. We have been referring to the descriptive aspects of statistical decision theory: its (accurate) predictions about how scientists *will* draw conclusions given evidence and prior probability. However statistical decision theory is prescriptive as well as descriptive. In addition to making predictions, it indicates how individuals facing uncertainty about true states of affairs *should* decide in order to maximize outcomes and minimize errors. According to Bayesian formulations, it is eminently *rational* to require more compelling evidence for low probability hypotheses than for those with high probability. In this way, the decision-maker is formally entitled (indeed required) to make use of previous knowledge, in drawing present conclusions. With classical procedures, by contrast, every conclusion is drawn as if in a vacuum, isolated from findings that have come before. It is true that the investigator typically endeavors to integrate his findings with prior evidence, but this process is informal, divorced from the statistical testing of hypotheses, rather than an integral part of it, as Bayesians would prefer.

Clearly, scientists are in the business of *deciding about* (rather than determining unequivocally) the validity of theories and hypotheses from accumulated evidence. Yet, as decision-makers, we have not taken advantage of the advances being made by mathematicians studying statistical decision theory. The proposed procedure incorporates their formulations and thus places our decisions on a more rational basis than the present.

An additional consideration is the unequal stature given by traditional procedures to Type I versus Type II (failing to reject a false null) errors. There is little philosophical basis to the argument that to embrace false statements is a more grievous error than failing to recognize the validity of true statements. While splitting the tails unequally would not go as far to restore the balance as the Bayesians might prefer, (they would wish us to specify quantitatively a different loss function for each problem) it is preferable on this dimension to classical procedures.

The procedure described here can be seen as a compromise not only between the use of the one- and the two-tailed tests. It is also a compromise between those who wish to maintain statements of significance and those who wish to abolish them. Bayesians, such as

Edwards, Lindman and Savage (1963) contend that beliefs change with each new datum, nonsignificant as well as significant. Similarly, Eysenck (1960) argues that the  $p$ -value for the test statistic has the lasting importance; whether or not  $p$  exceeds some arbitrary (.05) value and is called "significant" is irrelevant and has undesirable effects on the conduct of research. While the author has sympathy with these views, it is evident that their orientation has won few adherents among psychologists actively engaged in research. Manuscripts which contain statements of whether or not a finding is significant, without exact  $p$ -values, remain *de rigueur* for most researchers and editors. Perhaps, the proposed procedure, which retains the arbitrary  $\alpha$  level overall, but splits it between the two tails in accordance with a priori probability values, will be viewed as a realistic compromise.

## REFERENCES

- Bakan, D. The test of significance in psychological research. *Psychological Bulletin*, 1966, 66, 423-437.
- Burke, C. J. A brief note on one-tailed tests. *Psychological Bulletin*, 1953, 50, 384-387.
- Burke, C. J. Further remarks on one-tailed tests. *Psychological Bulletin*, 1954, 51, 587-590.
- Cohen, J. The statistical power of abnormal-social psychological research. *Journal of Abnormal and Social Psychology*, 1962, 65, 145-153.
- Edwards, W., Lindman, H., and Savage, L. J. Bayesian statistical inference for psychological research. *Psychological Review*, 1963, 70, 193-242.
- Eysenck, H. J. The concept of statistical significance and the controversy about one-tailed tests. *Psychological Review*, 1960, 67, 269-271.
- Goldfried, M. R. One-tailed tests and "unexpected" results. *Psychological Review*, 1959, 66, 79-80.
- Hick, W. E. A note on one-tailed and two-tailed tests. *Psychological Review*, 1952, 59, 316-318.
- Jones, L. V. Tests of hypotheses: One-sided *vs.* two-sided alternatives. *Psychological Bulletin*, 1952, 59, 43-46.
- Jones, L. V. A rejoinder on one-tailed tests. *Psychological Bulletin*, 1954, 51, 585-586.
- Kaiser, H. F. Directional statistical decisions. *Psychological Review*, 1960, 67, 160-167.
- Kimmel, H. D. Three criteria for the use of one-tailed tests. *Psychological Bulletin*, 1957, 54, 351-353.
- Marks, M. R. Two kinds of experiments distinguished in terms of statistical operations. *Psychological Review*, 1951, 58, 179-184.
- Marks, M. R. One- and two-tailed tests. *Psychological Review*, 1953, 60, 207-208.
- Meehl, P. E. Theory-testing in psychology and physics: A

methodological paradox. *Philosophy of Science*, 1967, 34, 103-115.

Overall, J. E. Classical statistical hypothesis testing within the context of Bayesian theory. *Psychological Bulletin*, 1969, 71, 285-295.

Schaffer, J. P. Directional statistical hypotheses and comparisons among means. *Psychological Bulletin*, 1972, 77, 195-197.

Tversky, A. and Kahneman, D. Belief in the law of small numbers. *Psychological Bulletin*, 1971, 76, 105-110.





## A NON-PARAMETRIC TEST FOR INCREASING TREND

MARTIN COOPER

Faculty of Education  
University of Ottawa

In this article, a new test for increasing or decreasing trend in a set of observations is presented. Rather than making specific paired comparisons, as in the Cox and Stuart tests for trend, the user of the new test makes comparisons of all possible pairs of observations. The advantage of this procedure is that, for a given number of observations, many paired-comparisons are made. This allows the researcher to examine increasing or decreasing trend when there is quite a small number of observations although, admittedly, the power of the test is low in such cases. A normal approximation is available for use with large samples.

PROBABLY the best known non-parametric tests for increasing or decreasing trend are those developed by Cox and Stuart. Of these, two are based on the binomial distribution. The other, which is for use with large samples only, used the normal deviate as the test statistic.

In each of the Cox-Stuart tests, the set of scores being tested for increasing or decreasing trend is divided into segments, corresponding scores in two of these segments being compared for relative magnitude. For the  $S_2$  test, the first half and the second half of the set of scores form the segments of interest, the middle score being dropped if there is an odd number of scores. For  $N$  scores, therefore, there are  $N/2$  comparisons at most. The  $S_3$  test uses the first and last thirds of the set of scores as the segments of interest, mid-scores being dropped (or dummy ones added) to make the effective number of scores equal to a multiple of 3. For this test, there are approximately  $N/3$  comparisons.

The first and last halves of the set of scores, which are used for the  $S_2$  test, are again used for the  $S_1$  test. For the latter test, however, the elements of the second segment are reversed in order.

Each of the Cox-Stuart tests thus involves the comparison of a series of pairs of corresponding scores. For each such pair, an indicator variable takes the value of unity if the later-occurring score is greater than the earlier-occurring, or zero if the opposite is true. The sum of the indicator values is the test statistic for the  $S_2$  and the  $S_3$  tests. The sum of weighted indicator values is the test statistic for the  $S_1$  test.

The purpose of the present article is to describe a new non-parametric test for increasing or decreasing trend. In this test, each score in a series is compared with each score which follows it. For a sample of  $N$  scores, therefore,  $\binom{N}{2}$  comparisons are made. In general, this is a considerably larger number of comparisons than would be made with any of the Cox-Stuart tests—a particularly important aspect when  $N$  is small. Since we are concerned with whether one score is greater than another, and not with the actual difference between them, scores may be discarded in favour of ranks.

Given the set of scores  $\{X_1, X_2, \dots, X_N\}$ , the problem is to test the null hypothesis that there is no monotonic increase in the set against the alternative hypothesis that monotonic increase exists. The ranks of the scores will be represented by  $\{r\} = \{r_1, r_2, \dots, r_i, \dots, r_j, \dots, r_N\}$ . Each pair of scores  $(X_i, X_j)$  is compared and an indicator variable  $a_{ij}$  takes the values

$$a_{ij} = 1 \quad \text{if } r_i - r_j < 0$$

$$a_{ij} = 0 \quad \text{if } r_i - r_j \nless 0$$

It is observed that the inclusion of tied observations in the category for which  $a_{ij} = 0$  will make the test more conservative. For each pair of scores,  $a_{ij}$  is a Bernoulli variable. The quantity  $\sum a_{ij}$ , summation being taken over all pairs, is thus binomially distributed, the parameters being  $\binom{N}{2}$  and  $\frac{1}{2}$ . The test involves the determination of whether the probability of the obtained value of  $S = \sum a_{ij}$ , and of all values which are more extreme, is greater than  $\alpha$ , the selected level of significance.

### *Numerical Example*

Is the set of scores  $\{11, 9, 14, 15, 13, 20\}$  monotonically increasing?

The corresponding set of ranks is  $\{2, 1, 4, 5, 3, 6\}$ . Calculation of the values of the indicator variable  $a_{ij}$  is performed below:

$i$	$j$	$r_j - r_i$	$a_{ij}$
1	2	1	0
1	3	-2	1
1	4	-3	1
1	5	-1	1
1	6	-4	1
2	3	-3	1
2	4	-4	1
2	5	-2	1
2	6	-5	1
3	1	-1	1
3	5	1	0
3	6	-2	1
4	5	2	0
4	6	-1	1
5	6	-3	1
			12

The sum of the values taken by the indicator variable is

$$s = \sum a_{ij} \\ = 12$$

The relevant binomial distribution is  $B(15, \frac{1}{2})$ , the first few terms of which are

$S$	$\binom{15}{S}(\frac{1}{2})^{15}$	
15	0.000	<div style="display: flex; align-items: center; justify-content: center;"> <div style="margin-right: 10px;">}</div> <div style="text-align: center;">0.017</div> <div style="margin-left: 10px;">}</div> </div>
14	0.000	
13	0.003	
12	0.014	
11	0.042	
10	0.092	0.059

The probability of  $S = 12$ , or of any  $S$  having a more extreme value, is 0.017. The null hypothesis is therefore rejected at the 5% level of

significance. The set of ranks is thus still sufficiently close to perfect ascending order to be regarded as monotonically increasing.

### *Large Samples*

The expectation of a random variable  $X$  having an associated binomial probability distribution  $B(n, p)$  is  $np$ ; the variance of  $X$  is  $np(1 - p)$ . As the value of  $n$  increases, the distribution  $B(n, p)$  approaches the normal distribution with mean  $E(X)$  and variance  $\text{var}(X)$ .

In cases where large numbers of pairs of scores are considered, therefore, a normal deviate may be used as the test statistic. Assuming, for illustrative purposes, that our numerical example constitutes a "large sample," the normal deviate is

$$\begin{aligned} z &= \frac{S - E(S)}{\sqrt{\text{var}(S)}} \\ &= \frac{12 - (15)(\frac{1}{2})}{\sqrt{(15)(\frac{1}{2})(\frac{1}{2})}} \\ &= 2.32 \end{aligned}$$

This value of  $z$  is greater than 1.64, and is thus significant at the 5% level.

### *Conclusion*

In conclusion, it is admitted that the comparison of large numbers of paired ranks is more tedious and time-consuming than the examination of only a few pairs. However, in an era when computers are readily available to most people who need them, this is not a serious drawback. On the other hand, the number of paired ranks greatly exceeds the number of observations. This not only allows us to place greater confidence in the data, there being more of them; it also allows us to use the test in cases when  $N$  is very small.

### REFERENCE

- Cox, D. R. and Stuart, A. Some quick tests for trend in location and dispersion. *Biometrika*, 1955, 42, 80-95.

## THE VARIANCES OF EMPIRICALLY DERIVED OPTION SCORING WEIGHTS<sup>1</sup>

GARY ECHTERNACHT  
Educational Testing Service

Estimates for the variances of empirically determined scoring weights are given. It is also shown that test item writers should write distractors that discriminate on the criterion variable when this type of scoring is used.

IN recent years, the developers of large-scale testing operations have shown an increasing interest in reducing the length of time examinees are required to spend on a given test. Reducing the test administration time would both reduce the cost of developing the test forms, as fewer items would be required, and allow time for additional tests to be administered. This thinking has characterized many of the test programs administered at Educational Testing Service, and, most likely, at other testing establishments. Researchers have thus sought new scoring methods that would result in increases in reliability due solely to the scoring system used. Thus, test length could be reduced, and a previous standard of reliability maintained.

One such scoring method that has proven successful in reliability studies is that of empirically deriving scoring weights (Davis and Fifer, 1959; Echternacht, 1973; Hendrickson, 1971; Reilly and Jackson, 1972; Strong, 1943). If empirically derived scoring weights were to be adopted by such large-scale testing programs as the College Entrance Examination Board, the Graduate Record Examinations, the Law School Admission Test, and other programs, one problem that would have to be faced is that of determining the variances of the derived weights and the implications these variances have for developing test

---

<sup>1</sup> This research was supported by the Graduate Record Examinations Board.



items. This is necessary on repeated occasions, and the scoring weights would only be developed on the initial administration. Since some examinees would not be included in the initial scoring run, the problem of scoring weight variance exists. Also, by knowing this variance, the minimum number of examinees needed to develop the weights, subject to a specified level of precision, can be determined.

There are a number of methods that can be used for deriving the weights. The method that will be discussed here is that used by Echternacht (1973), which is actually the method used by Reilly and Jackson (1972) with no iterations. Briefly, the method consists of assigning the average criterion score of those selecting a given option. The criterion variable is standardized, so that its mean is zero and variance is one. The criterion that is usually used is the score on the remaining items that make up the test although this is certainly not a necessary criterion.

Consider a population of  $N$  people who will take a given test at one point in time. Assume further that a simple random sample of  $n$  people from the population take the test for the purpose of determining scoring weights. Although this is not exactly true in an operational setting, it does provide a useful approximation to reality. Consider one item for that test. The scoring weight assigned to the  $i$ th option of this item is

$$\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$$

where  $n_i$  represents the number of people responding with the  $i$ th option and  $y_{ij}$  represents the criterion score for the  $j$ th person choosing the  $i$ th option. In weighting options, the omit category is considered another option and a weight is also derived. Since the criterion variable is assumed to be standardized,

$$\sum_{i=1}^c n_i \bar{y}_i / n = \bar{y}_{..} = 0$$

where

$$n = \sum_{i=1}^c n_i,$$

the number of people responding to the item with one of the  $c$  possible options. Using the standard result for the variance of a mean obtained by simple random sampling from a finite population, the variance of the  $i$ th option weight thus becomes

$$(1/n_i - 1/N_i) S_i^2$$

where

$$S_i^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (Y_{ij} - \bar{Y}_i)^2.$$

$N_i$  indicates the number of examinees in the population responding with option  $i$ . The problem becomes one of estimating  $S_i^2$ . This is done by using the unbiased estimate

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2.$$

Such estimates of  $S_i^2$  would presumably be obtained through pretesting of the item.

Suppose the whole population of  $N$  examinees is used for the purpose of determining scoring weights, and the method previously described is used.

Now,

$$S^2 = \frac{1}{N - 1} \sum_i^c \sum_{j=1}^{N_i} (Y_{ij} - \bar{Y}_i)^2 = 1, \quad \text{and} \quad \bar{Y}_i = 0$$

where  $c$  indicates the number of response options. From the standard algebraic identity for the analysis of variance, with

$$N = \sum_{i=1}^c N_i,$$

$$\begin{aligned} (N - 1)S^2 &= (N - 1) = \sum_{i=1}^c \sum_{j=1}^{N_i} (Y_{ij} - \bar{Y}_i)^2 \\ &= \sum_{i=1}^c N_i (\bar{Y}_i - \bar{Y}_i)^2 + \sum_{i=1}^c \sum_{j=1}^{N_i} (Y_{ij} - \bar{Y}_i)^2 \\ &= \sum_{i=1}^c N_i \bar{Y}_i^2 + \sum_{i=1}^c (N_i - 1) S_i^2. \end{aligned} \quad (1)$$

If the  $1/N$  is negligible (1) may be written as

$$1 = \sum_{i=1}^c W_i \bar{Y}_i^2 + \sum_{i=1}^c W_i S_i^2, \quad (2)$$

where

$$W_i = N_i/N$$

so that

$$\sum_{i=1}^c W_i (1 - \bar{Y}_i^2) = \sum_{i=1}^c W_i S_i^2$$

and

$$\sum_{i=1}^c W_i(1 - \bar{Y}_i)^2 = \sum_{i=1}^c W_i S_i^2 = \bar{S}^2, \quad (3)$$

which indicates that the  $S_i^2$  are not independent for all  $c$  categories.

In obtaining empirically derived scoring weights it is, of course, desirable to have the variance of the resulting weights be of a minimum. If a large enough pool of examinees are tested in the initial test administration so that the  $n_i$  are all large for each item, the variances will likely be small. This is not always the case, though, and it does not tell the item writer anything about how he should write the items to help insure that a small variance results. The item writer can have some influence over both the  $n_i$  and the  $S_i^2$ . By increasing the  $n_i$  and decreasing the  $S_i^2$  and the  $i$ th option weight's variance will decrease. But, the  $n_i$  and  $S_i^2$  are not independent for a given item. Therefore, it seems reasonable to consider minimizing  $\bar{S}^2$  and the implications this minimization has for item writers. One can see that  $\bar{S}^2$  can be minimized by making the between options sum of squares,  $\sum_{i=1}^c N_i \bar{Y}_i^2$ , a maximum.

Although it is recognized that the following discussion is somewhat esoteric for the item writer and the conditions presented very unrealistic, the discussion following is an attempt to demonstrate some of the basic principles that should be used in minimizing  $\bar{S}^2$ . In maximizing  $Q = \sum_{i=1}^c N_i \bar{Y}_i^2$ , a few things need to be noted. In the case where  $c = 2$ , it can be easily shown that  $Q$  attains a minimum when  $\sum_{j=1}^{N_i} Y_{ij} = 0$ , or when each category mean equals the overall mean. Also, if  $\sum_{j=1}^{N_i} Y_{ij}$  can be considered given and  $Q$  a function of only the  $N_i$ 's,  $Q$  is minimized when  $N_i = N/2$ . Since we are considering a finite population, a maximum value of  $Q$  is obtained when all positive  $Y_{ij}$  are found in one category and all negative  $Y_{ij}$  in the other. The zero values of  $Y_{ij}$  are placed in the category with the largest  $N_i$ .

In cases where  $c > 2$ , it can be shown that  $Q$  is minimized when  $\sum_{j=1}^{N_i} Y_{ij} = 0$  for each  $i$ , or if the sums,  $\sum_{j=1}^{N_i} Y_{ij}$ , are considered fixed, when the  $N_i$  are proportional to  $|\sum_{j=1}^{N_i} Y_{ij}|$ . Maximum values can be obtained only when the criterion values can be partitioned into nonoverlapping regions, with each region corresponding to a group of people responding with a particular distractor. In topological terms these regions are termed "connected" regions, and their union consists of the entire criterion variable space. This is also the case where each distractor can be used to place the individual responding with that distractor in a categorization of the criterion.

In practice though, it is impossible for an item writer to write items with the property previously noted. The item writer can structure dis-

tractors in such a way that examinees of differing ability levels respond to different distractors. Such a practice would tend to approximate the condition mentioned previously, assuming that ability and the criterion are related, and allow  $Q$  to be maximized as much as is practical. The procedure of "facet design" as set forth by Guttman (see Elizur, 1970) is one method that might be used to so structure the distractors. In examining the results of item pretesting, the quantity  $Q$  should also be taken into consideration in making the decision of whether or not to include a given item as part of a test that will be scored using empirically derived option weights.

## REFERENCES

- Davis, F. B. and Fifer, G. The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1959, 19, 159-170.
- Echternacht, G. J. A comparison on various item option weighting schemes. Research Bulletin 73-6. Princeton, N. J.: Educational Testing Service, 1973.
- Elizur, D. *Adapting to innovation*. Jerusalem, Israel: Jerusalem Academic Press, 1970.
- Hendrickson, G. F. The effect of differential option weighting on multiple-choice objective tests. *Journal of Educational Measurement*, 1971, 8, 291-296.
- Reilly, R. R. and Jackson, R. Effects of empirical option weighting on reliability and validity of the GRE. Research Bulletin 72-38. Princeton, N. J.: Educational Testing Service, 1972.
- Strong, E. K., Jr. *Vocational interests for men and women*. Stanford, Calif.: Stanford University Press, 1943.





## THE APPLICATION OF DISCRIMINANT FUNCTION ANALYSIS TO CORRELATED SAMPLES

G. FRANK LAWLIS  
Texas Tech University

ARTHUR B. SWENEY  
Wichita State University

A method of applying the discriminant function to correlated samples was discussed for the researcher interested in utilizing multivariate statistics to pre-post data. By finding the linear combination that maximizes differences in the groups, the  $t$ -statistic can be computed for correlated samples.

FOR years researchers in the behavioral sciences have been attempting to demonstrate the general effectiveness of therapeutic treatments. Although there are several designs from which to infer such influences, two methods come readily to mind. From one method one can compare separate groups that have had, or have been assumed to have had, differential treatments, i.e., the cross-sectional design. The other design involves the comparison to the same group over a series of time segments, i.e., the longitudinal approach.

Statistical methods have been devised to determine if treatment outcomes significantly differ with respect to a dependent variable, i.e., analysis of variance for block effects,  $t$ -tests for correlated samples, etc. However, these methods are utilized for only one dependent variable.

For each variable tested, the probability of finding significance by chance alone increases geometrically. Consequently, researchers have to make decisions as to what dependent variables appear to be critical to their particular model of research. Theory building becomes a process of inductive reasoning variable by variable testing.

The measurement of the effectiveness of a treatment utilizing a combination of variables can be facilitated through the use of discriminant analysis as the most appropriate statistical method since the procedure makes it possible to maximize the group difference by the most efficient combination of variables. Discriminant analysis is virtually always applied to discrete groups, and thus limited to cross-sectional studies. A problem arises when a researcher wishes to show change of one group over time with a combination of variables. In other words, the discriminant function has not been applied to longitudinal research. The purpose of this paper is to consider an application of the discriminant analysis to correlated or identical groups measured through time.

### *Method*

Before the proposed technique is discussed, a simple explanation of the principle of the discriminant function should be presented. Consider the simplest case in which a researcher wished to combine two variables such as ego strength (A) and motivation (B) to discriminate between two discrete groups, successful therapy and nonsuccessful therapy cases, as determined by the researcher. If the researcher can use the scales as a coordinate system for a two-dimension model, those points can be represented in two-dimensional space.

As seen in Figure 1, those points can be transformed to another linear scale C in which the group differences are maximized. That linear scale C will pass through the origin (0, 0) and the coordinates that would be the respective weights of the variables. In our example, the weights that maximized group difference were 2.00 A (ego strength score) and -1.00 B (motivation). Each value can be weighted, summed and represented on the vector as a single point.

The strategy is to determine the eigenvector that will provide the greatest separation between the two groups such that the between variance is maximized.

$$\text{Max variance} = \frac{\text{Sum of Squares between}}{\text{Sum of Squares within}}$$

There can be  $n$ th dimensional space with  $n$  number of variables, and there is no room to discuss the calculus of determining the respective weights in which orthogonal relationship could be assumed. For those researchers interested in these procedures, please refer to Tatsuoka (1970) for a more sophisticated and thorough explanation of these more complex computations.

The primary consideration is that there is a possibility of represent-

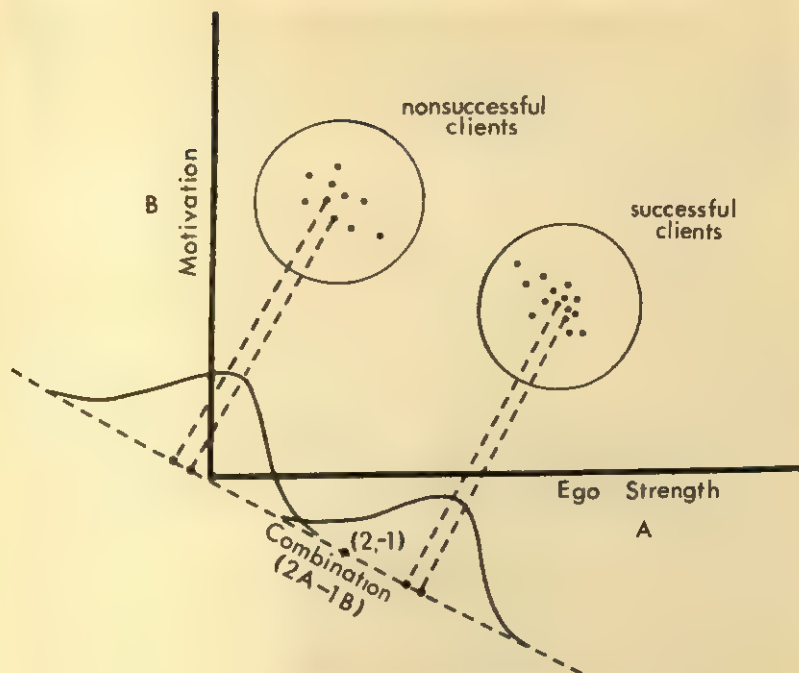


Figure 1. Data of two groups represented in two-dimensional space.

ing the combination of variables in linear form maximizing group differences. Therefore, the final variable scores have the same property of occurring randomly as any of their sub-parts. As such, it can be used in reference to the *t*-distribution.

The major problem in applying such a technique to dependent samples is the assumption of interdependence between groups. That is, in most cases there is a positive correlation between pre- and post-scores. Therefore, one must make them statistically independent by subtracting the variance attributable to pre-test score from the post-score as follows:

$$Z_{\text{post}}^1 = Z_{\text{post}} - r_{\text{pp}} Z_{\text{pre}}$$

$Z_{\text{post}}^1$  = partialled score on post-test after variance explainable by pre-test is removed

$Z_{\text{post}}$  = standard scores on post-test

$Z_{\text{pre}}$  = standard scores on pre-test

$r_{\text{pp}}$  = PM correlation between pre- and post-scores

It is then a simple matter to apply the *t*-test for *correlated* samples to that linear combination vector (Ferguson, 1959) in showing changes over time for a combination of variables. Moreover, by the value of the maximizing weights, a researcher could determine the most relevant variables to that change.

### Discussion and Implications

Psychological research has frequently been directed toward determining the effects of therapeutic intervention, yet change is difficult to attribute to any one variable, whether it is depression, disruptive behavior, or the disability. Moreover, error rate prohibits the statistical testing of a wide range of important variables analyzing each one by one. With the utilization of a combination of variables or indices of variables, such as behavior ratings, test scores, etc., the researcher may more easily demonstrate an outcome. As an illustration, consider that the researcher found the combination of ego strength (A) and motivation scores (B) with the weights of 2.0 and -1.0, respectively, maximized the pre-test scores from the post-test scores significantly (See Table 1). Finding a significant change from pre- to post-testing, he could make the inference that his treatment ap-

TABLE 1  
Illustration Problem

Ss	Pre-Score*			Post-Score*			D	D
	Ego Strength	Motivation	Combination**	Ego Strength	Motivation	Combination**		
1	5	4	6	5	4	6	0	0
2	4	5	3	6	8	4	-1	1
3	5	4	6	7	6	8	-2	4
4	3	3	3	8	4	12	-9	81
5	2	4	0	5	4	6	-6	36
6	5	5	5	6	5	7	-2	4
7	3	5	1	7	10	4	-3	9
8	4	6	2	6	8	4	-2	4
9	6	6	6	8	8	8	-2	4
10	5	5	5	6	5	7	-2	24
							-29	147

\* Stanines ( $\bar{X} = 5$ , S.D. = 2)

\*\* Weights Ego Strength = 2 Motivation

$$t = \frac{-2.9}{\sqrt{\frac{147}{10} - (-2.9)^2/9}} = -3.4 \quad p < .05.$$

peared to be effective in making change for this combination of variables, ego strength being the more relevant.

Perhaps this application of the discriminant analysis is overly simplistic, but the belief that change occurs in only one variable at a time is a more simplistic concept which has probably drastically limited complex theory building in psychology. Obviously, replication of findings using identical weights would show validation of the results, but trends could be determined by statistical confirmation. At least a method of deductive hypothesis testing could be realized.

## REFERENCES

- Ferguson, G. *Statistical analysis in psychology and education*. New York: McGraw-Hill, 1959.
- Tatsuoka, M. *Selected topics in advanced statistics*, No. 6. Champaign: Institute of Personality and Ability Testing, 1970.





## A COMPARISON OF VARIABLE CONFIGURATIONS ACROSS SCALE LENGTHS: AN EMPIRICAL STUDY<sup>1</sup>

HOWARD G. SCHUTZ AND MARGARET H. RUCKER  
University of California, Davis

Data from 2-, 3-, 6-, and 7-point rating scales were analyzed to determine whether scale length affected response patterns. The results of this study indicate that data configurations are relatively invariant with changes in number of scale points.

IN educational and psychological research, rating scales are often used to collect data. An important question in constructing such scales is how many response categories to provide. This problem has been approached in several ways. Bendig (1954) and Komorita and Graham (1965) studied the reliability of information obtained using scales of different lengths. Their research indicates that reliability, at least of the sum of a set of homogeneous rating scales, is independent of number of response categories. Research by Matell and Jacoby (1971) supported these results and also revealed that both stability and validity of cumulative scores from Likert-type items are independent of the number of scale points. Finn (1972), however, found that for ratings on a single factor, ratings on a 9-point scale were less reliable than ratings on shorter scales. Other aspects of the problem studied by Matell and Jacoby (1972) include the effects of testing time and scale properties. This research revealed that proportion of scale used was independent of number of scale points (excluding 2- and 3-point scales) but mean testing time increased and usage of an "uncertain" category decreased as number of scale steps increased. Green and Rao (1970), using numerical simulation, investigated the effect of scale length on

<sup>1</sup> The authors wish to thank Professor G. F. Russell for his development of the computer program used to calculate means, cross products, and similarity statistics used in this research.

ability to recover the original data configuration. Their research indicates that one should use at least six response categories—response degradation to 3-point or 2-point scales results in poor recovery of the original configuration. The present study was designed to extend these results by comparing factors produced from empirical data.

### *Research Design*

As part of an ongoing food attitude research project, the authors developed four forms of a food-use questionnaire to investigate the effects of scale length on response patterns. These questionnaires were identical except for number of scale points. The questionnaires called for rating the appropriateness of ten different foods in ten different situations. The ten foods were utilized as the variables in this study. An example of these questionnaires is given in Figure 1.

It was decided to include 2-, 3-, 6-, and 7-point scales in this study. These numbers were chosen to make it possible to compare short and the more commonly used longer scales, with and without midpoints. All four scales were anchored at the ends with the terms "appropriate" and "inappropriate."

Subjects were 60 male and 60 female students enrolled in a history course at a large western university. Fifteen male and fifteen female students were randomly assigned to complete each of the four forms of the questionnaire.

### *Analyses and Results*

For each group, a mean rating for each food-use combination was computed. Raw cross-products computed from the mean ratings were factor analyzed to produce clusters, as suggested by Nunnally (1967, p. 381). The Biomedical computer program used for these analyses performs a principal component solution and an orthogonal rotation of the factor matrix.

For each of the scales, three factors accounted for over 97% of the variance: .993 for the 2-point, .988 for the 3-point, .982 for the 6-point, and .979 for the 7-point.

The factor loadings were then examined for differences between the scales. Since the size of cross-product factor loadings varies with scale size, the loadings were converted to proportions to facilitate this inspection. These loadings are presented in Table 1.

To obtain a measure of variability of the factor loadings that was comparable across scales, the standard deviations of the factor loadings were converted to coefficients of variation. The resulting figures,



TABLE 1  
Factor Loadings as Proportions<sup>a</sup>

Foods <sup>b</sup>	Factor 1				Factor 2				Factor 3			
	2-point scale	3-point scale	6-point scale	7-point scale	2-point scale	3-point scale	6-point scale	7-point scale	2-point scale	3-point scale	6-point scale	7-point scale
1	.099	.099	.097	.106	.120	.128	.127	.141	.090	.088	.083	.090
2	.124	.133	.137	.148	.085	.071	.069	.066	.098	.100	.102	.100
3	.086	.075	.067	.073	.092	.094	.097	.089	.108	.115	.111	.113
4	.102	.102	.100	.103	.111	.113	.121	.137	.080	.075	.072	.078
5	.107	.112	.118	.121	.093	.101	.102	.097	.104	.089	.099	.091
6	.078	.067	.067	.063	.127	.134	.136	.144	.104	.101	.106	.113
7	.103	.103	.107	.098	.088	.085	.074	.063	.106	.127	.121	.118
8	.112	.124	.132	.124	.115	.119	.129	.126	.111	.116	.130	.130
9	.075	.071	.059	.051	.080	.078	.066	.063	.115	.118	.119	.118
10	.116	.113	.117	.114	.089	.077	.079	.074	.084	.073	.057	.048

<sup>a</sup> Presented as proportions to obviate effect of scale size on loading size

<sup>b</sup> Given in order as they appear in questionnaire, Figure 1.



TABLE 2  
Coefficients of Variation of Cross-Product Factor Loadings

	Factor 1	Factor 2	Factor 3
Two-point scale	.1624	.1665	.1171
Three-point scale	.2244	.2260	.1878
Six-point scale	.2773	.2710	.2300
Seven-point scale	.2984	.3387	.2405

1967). The only discernible trend in the results of these analyses was that the correlation between the 2-point and 7-point scale factor loadings was slightly lower than the other correlations, for all three factors. However, the coefficients were all so uniformly high—.98 or higher for all comparisons—that the importance of this trend is questionable.

A measure of the distance between profiles,  $D$ , was computed between means for all pairs of foods (Nunnally, 1967). The  $D$  values for each scale were divided by the maximum possible  $D$  for that scale, and these numbers were then subtracted from 1 so that the final values would range from 0 to 1 with 1 being identity.

A Kruskal (1964) nonmetric multidimensional analysis was computed on the similarity data from each group. On the basis of the factor analyses results, three dimensions were selected to be fitted in the Kruskal analyses. The resulting stress values were .042 for the 2-point, .028 for the 3-point, .018 for the 6-point, and .018 for the 7-point scale. There is some tendency for stress to decrease as number of response categories increases through 6-points. However, there is no difference between the 6- and 7-point stress values and all of the stress values are low.

Responses for each scale were then subdivided into male and female groups. Analyses of these subgroups produced no readily evident trends, either between males and females or across scales. However, different results may be produced with larger sample sizes. It appears that one does not get a stable factor structure with only 15 cases.

### Conclusion

Within the limits of this study, it is concluded that number of available response categories, at least within the 2- to 7-point range, does not materially affect the cognitive structure derived from responses to that scale.

### REFERENCES

- Bendig, A. W. Reliability and the number of rating scale categories. *Journal of Applied Psychology*, 1954, 38, 38-40.

- Finn, R. H. Effects of some variations in rating scale characteristics on the means and reliabilities of ratings. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1972, 32, 255-265.
- Green, P. E. and Rao, V. R. Rating scales and information recovery—How many scales and response categories to use? *Journal of Marketing*, 1970, 34, 33-39.
- Harman, H. M. *Modern factor analysis* (2nd ed.). Chicago: University of Chicago Press, 1967.
- Komorita, S. S. and Graham, W. K. Number of scale points and the reliability of scales. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1965, 25, 987-995.
- Kruskal, J. B. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 1964, 29, 115-129.
- Matell, M. S. and Jacoby, J. Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and validity. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1971, 31, 657-674.
- Matell, M. S. and Jacoby, J. Is there an optimal number of alternatives for Likert-scale items? Effects of testing time and scale properties. *Journal of Applied Psychology*, 1972, 56, 506-509.
- Nunnally, J. C. *Psychometric theory*. New York: McGraw-Hill, 1967.
- Tucker, L. R. A method for synthesis of factor analysis studies. Personnel Research Section report, No. 984. Washington, D. C.: Department of the Army, 1951.

## AN INVESTIGATION OF THE RASCH SIMPLE LOGISTIC MODEL: SAMPLE FREE ITEM AND TEST CALIBRATION<sup>1</sup>

HOWARD E. A. TINSLEY<sup>2</sup> AND RENÉ V. DAWIS

University of Minnesota

This research investigated the use of the Rasch simple logistic model in item and test calibration. Tests employing word, picture, symbol, and number analogies were administered to high school students, college students, civil service clerical employees, and clients of the Minnesota Division of Vocational Rehabilitation. The results indicated that Rasch item easiness ratios and  $z$  item difficulty ratios were invariant with respect to the ability of the calibrating sample when an adequate sample was employed and the test design did not incorporate biasing factors. The invariance of the Rasch item easiness ratios was shown to be related to the goodness-of-fit of the items to the Rasch model in that the deletion of items with low Rasch probabilities increased the invariance of the Rasch item easiness ratios. The estimation of the amount of ability indicated by the raw scores on a test was also shown to be invariant with respect to the ability of the calibrating sample for tests of 25 or more items, even when samples of fewer than 100 subjects were studied.

OVER 20 years ago Gulliksen (1950) remarked that the discovery of item parameters which would remain stable as the item analysis group changed would constitute a significant contribution to item analysis theory. More recently Lord and Novick (1968) have expressed

<sup>1</sup> This research is based on the first author's doctoral dissertation for which the second author was a principal advisor. A condensed version of the portion of this paper dealing with test calibration was presented at the Western Psychological Association Convention, Portland, Oregon, April, 1972. This research was supported by grant number N00014-68-A-0141-0003 from The Office of Naval Research to The Center for the Study of Organizational Performance and Human Effectiveness, by a grant from the Minnesota Research Coordinating Unit for Vocational Education, and by a grant from the University of Minnesota Department of Psychology.

<sup>2</sup> Now in the Department of Psychology, Southern Illinois University, Carbondale, Illinois 62901.

a similar opinion. Within the framework of classical test theory a number of indices of item difficulty have been suggested which might possess this property. A normal curve transformation of  $P$  values to  $z$  values, frequently referred to as Thurstone's method of absolute scaling, has been suggested by several authors (Bliss, 1929; Guilford, 1954; Horst, 1933; Thorndike, Bergman, Cobb, and Woodyard, 1926; Thurstone, 1925, 1947). A second method commonly suggested for obtaining invariant item difficulty parameters, the limen method, has been described by Bliss (1929), Thorndike et al. (1926), and Tucker (1953). Modifications of the limen method have been discussed by Gulliksen (1950) and Richardson (1936). Both the method of absolute scaling and the limen method require the assumption of a normal distribution for the ability under consideration. Although both methods were first described 50 years ago, neither apparently has been investigated systematically.

More recently, Rasch has introduced a "latent trait" model which purportedly makes possible sample-free item and test calibration (Rasch, 1960, 1961, 1966a, 1966b). A major advantage of the model is its "objectivity," i.e., the model allows the computation of item and test parameters from any sample of subjects since the estimation of the parameters is independent of the distribution of ability in the calibrating sample. Schmidt (1970) has presented a proof that the Rasch model is the only model to produce objectivity. The purpose of this study was to investigate the objectivity of the Rasch model in item and test calibration.

The Rasch model makes the following assumptions (Anderson, Kearney, and Everett, 1968; Brooks, 1965; Sitgreaves, 1963):

1. Items are scored dichotomously,
2. Speed does not influence the probability of a correct response,
3. Given the parameters for item easiness ( $e$ ) and subject ability ( $a$ ), all responses on a test are stochastically independent, and
4. The probability of a correct response by individual  $i$  to item  $j$  is a function of the ratio  $a_i/e_j$ .

This last assumption excludes guessing and variations in item discrimination as factors which affect the probability of a correct response. The effects of violating this assumption have been studied by Brink (1971) and Panchapakesan (1969).

Only three investigations of item-calibration using the Rasch model have been reported in the literature. Rasch (1960) used data from four subtests of the Danish Military Group Intelligence Test BPP which was given to 1094 Danish military recruits. He found the data fit his model for subtests  $N$  (a test of finding the next term in a numerical sequence) and  $L$  (a test similar to Raven's Progressive Matrices, but with groups

of letters instead of geometric figures). The model was inadequate to explain performance on subtests *F* (a test in which geometric shapes are to be decomposed into parts) and *V* (a test of verbal analogies). Rasch had used restrictive time limits with subtests *F* and *V*, however. When the time factor was controlled, the data for these subtests also fit his model (Rasch, 1966a).

Brooks (1965) wanted to determine if intelligence test data obtained from American public school children would fit the Rasch model. Samples of eighth graders and tenth graders in Iowa Public Schools (part of the standardization sample for the 1964 Lorge-Thorndike Intelligence Test) were employed in this study. Of the 243 items tested, 177 (72.8%) fit the Rasch model. Brooks (1965) also investigated the invariance of item easiness ratios and concluded that Rasch item easiness ratios are invariant with respect to the ability of the calibrating sample.

Anderson et al. (1968) investigated the hypotheses that Rasch item easiness estimates are independent of the ability of the calibrating sample, and that Rasch item easiness estimates are more stable when only items which fit the Rasch model are considered. The test used was the 45-item spiral omnibus intelligence test for screening applicants to the Australian Army or Royal Australian Navy. Samples of 608 recruit applicants to the Citizen Military Force (CMF), and 874 recruit applicants to the Royal Australian Navy (RAN) were studied. Twelve items were deleted for zero or 100% correct responses. For the CMF sample 30 items (91%) fit the Rasch model at the .01 level of confidence, and 25 items (76%) fit the Rasch model at the more stringent .05 level of confidence. (The level of confidence represents the probability of obtaining the observed pattern of responses, assuming the Rasch model is adequate to explain performance on the item.) For the RAN sample the corresponding findings were 22 items (67%) and 16 items (48%). The authors computed the product-moment correlation between the item easiness estimates obtained from the CMF and RAN samples. The authors concluded from the correlation of .958 (based on 33 items) that the item easiness ratios were independent of the ability of the samples upon which they were computed. When those items that failed to fit the Rasch model at the .05 level were deleted, a correlation increased to .990.

Calibrating a test using the Rasch model results in a logarithmic ability estimate being assigned to every possible raw score from 1 to  $K-1$  ( $K$  = number of items). This estimate indicates the amount of ability required to achieve that raw score. A comparison of the ability estimates assigned to a given raw score by two samples of different



ability should indicate the degree to which the Rasch model calibrates a test independently of the ability level of the calibration sample.

Wright (1967) reports one such investigation based on the responses of 976 beginning law students to 48 reading comprehension items on the Law School Admission Test. To obtain samples of different ability Wright selected two contrasting groups from his total sample. The "dumb group" included the 325 students who did poorest on the test, with a top score of 23. The "smart group" included the 303 students with the highest scores, their lowest score being 33. Wright compared the similarity between the two sets of Rasch ability estimates and the two sets of percentile ranks. He concluded that the Rasch model leads to sample-free test calibration while the "traditional" method does not.

Anderson et al. (1968) studied the invariance of Rasch ability estimates. They correlated the ability estimates obtained from the CMF and RAN samples. The resulting product-moment correlation of .992 was interpreted as evidence that the ability estimate assigned to a score on a test is independent of the distribution of ability in the calibrating sample. However, it is doubtful that the two samples actually differed in ability.

This paper examines the application of the Rasch model to analogy test items. The following hypotheses were investigated:

1. Rasch item easiness ratios are invariant with respect to the ability level of the calibrating sample.
2. The higher the probabilities that the individual items fit the Rasch model, the more invariant the item easiness ratios are with respect to the ability level of the calibrating sample.
3. Rasch ability estimates, assigned in the calibration of a test, are invariant with respect to the ability level of the calibrating sample.

To provide a base line against which the invariance of the Rasch item easiness ratios can be compared, a conventional item easiness parameter (the  $z$  item difficulty index) was also calculated and submitted to similar tests.

### *Method*

#### *Selection of Item Format*

Spearman's "g" or general mental ability seems to be represented in almost all the major intelligence tests in use today. Helmstadter (1964) points out that tests dealing with abstract relationships (such as verbal, numerical, or symbolic analogies) come closest to representing what is

meant by "g." For this reason, the analogy format was selected for study in this research. Guilford (1959) suggests that there are several different methods of asking analogy questions, i.e. figurally, symbolically, semantically, and behaviorally, depending upon the type of material used to present the question. To make the present study as general as possible, it was decided to study figural (picture), symbolic (number and symbol), and semantic (word) test items. Two types of symbolic material were used because of the intrinsic differences in the two and because Guilford (1966) has reported the discovery of more than one factor in some cells in his Structure-of-Intellect.

### *Subjects*

Data were obtained from four samples of subjects: college students enrolled in an introductory psychology class at the University of Minnesota; high school students enrolled in two suburban Twin Cities high schools; civil service clerical employees of the city of Minneapolis; and clients of the Minnesota State Division of Vocational Rehabilitation (DVR).

The samples were similar in terms of race, religion, and sex. The high school and college students were younger than the DVR clients and civil service employees, had fewer marital obligations, were better educated, came from homes which had higher family incomes, had better educated mothers, and had fathers employed in higher level occupations. In comparison with the high school and college students, the civil service employees were older, had lower family incomes, and were far more likely to be married and have children. The DVR clients constituted the most heterogeneous sample in many respects but were less well educated and had lower family incomes than did the high school and college students.

### *Instruments*

Four tests were used with the college and high school students: a 60-item word analogy test, a 60-item number analogy test, a 50-item picture analogy test, and a 40-item symbol analogy test. (For a discussion of the test construction process, see Tinsley, 1972.) A 25-item word analogy test was used with the DVR clients, while a 30-item picture and a 30-item word analogy tests were administered to the Minneapolis civil service employees. (These word and picture analogies had been selected in an unusual manner. The picture items were selected from picture items surviving an iterative item analysis procedure [see Tinsley, 1972]. The word analogies were then con-

structed from the picture analogies by substituting the word for the object in the picture.)

Item responses were scored and submitted to analysis, using a computer program written by Wright and Panchapakesan (1969, 1970) and modified by Bart, Lele, and Rosse (1970).

The first question of interest was whether the use of the Rasch model leads to item easiness ratios that are invariant with respect to the ability of the calibrating sample. Ten "two-sample" comparisons were made in this study (see Table 1). In each case a set of analogy items was completed by two samples differing in ability. The two sets of data were then independently submitted to item analysis. The product-moment correlation was calculated between the two sets of Rasch item easiness estimates and, for comparison purposes, between the two sets of  $z$  item difficulty estimates.

The relationship between the "goodness-of-fit" of the item and its invariance was also studied. The Rasch item easiness estimates obtained for the two samples were correlated, first for all items, then for the remaining items after first eliminating those items that failed to fit the Rasch model for both samples at, respectively, the .01, .05, .10, .25, .30, .35, and .40 levels of confidence. A similar procedure was employed in investigating the relationship between the invariance of the  $z$  item difficulty estimate and the "goodness-of-fit" of the  $P$  value. The criterion levels used for this index were  $.20 \leq P \leq .80$ ,  $.30 \leq P \leq .70$ , and  $.40 \leq P \leq .60$ . In both cases, the hypothesis was that the between-sample correlation would increase as the criterion became more stringent.

Finally, the invariance of the ability estimates computed for each

TABLE 1  
*Comparisons Made in Testing Invariance of Rasch Item Easiness Ratios*

Comparison Code Number	Sample 1	Sample		N	Analogy Items Type Numbers
		N	Sample 2		
I	College	630	High School	319	Word 60
II	College	630	DVR Clients	89	25
III	High School	319	DVR Clients	89	25
IV	College	276	Civil Service	269	30
V	College	492	High School	120	Picture 50
VI	College	492	Civil Service	269	25
VII	High School	120	Civil Service	269	25
VIII	College	276	Civil Service	269	30
IX	College	492	High School	145	Number 60
X	College	630	High School	308	Symbol 40

raw score was investigated in each of the ten comparisons by computing the product-moment correlation between the two sets of independently obtained ability estimates.

## *Results*

### *Item Calibration*

Ten sets of data were relevant to an investigation of the invariance for Rasch item easiness and  $z$  item difficulty ratios. Tables 2 and 3 show the results of analysis of these data. For all items, in all but one comparison, the correlation between independent estimates of Rasch item easiness differed no more than one point from the correlation between independent estimates of the  $z$  item difficulty index.

Four tests of the invariance of these parameters were performed with word analogies. The Rasch item easiness estimates obtained from college students correlated highly with those obtained from high school students (.95, comparison I) and civil service employees (.91, comparison IV). At the other extreme, the Rasch item easiness estimates obtained from DVR clients had near zero correlations with those obtained from college and high school students (comparisons II and III).

Four tests of the invariance of the item parameters were also conducted with picture analogies. The Rasch item easiness estimates obtained on the 50-item and 30-item picture analogy tests showed high correlations (comparisons V and VIII), while those obtained on the 25-item test showed low correlations (comparisons VI and VII).

Item parameters obtained on the 60-item number analogy and the 40-item symbol analogy tests yielded high correlations (comparisons IX and X).

The above results indicate the degree to which the item parameters are invariant when the analysis is performed on all items in the test. The Rasch model, however, cannot be expected to hold for items which do not fit the model. For this reason the relationship between the invariance of the item parameters and the "goodness-of-fit" of the item was investigated.

Elimination of items which did not fit the Rasch model resulted in some increase in the correlation between Rasch item easiness estimates, but the results did not follow a simple pattern. Only comparison VIII (between college students and civil service clerical employees on 30 picture analogies) showed a steady decrease in invariance as items with lower Rasch probabilities were removed. In contrast, comparison VII (between high school students and civil ser-

TABLE 2  
Correlation of Rasch Item Easiness Estimates

Minimum Rasch Probability	Comparison and Type of Analogies										Symbol X	
	I	II	Word	III	IV	V	VI	Picture	VII	VIII		Number IX
All Items	.95	-.08		.05	.91	.97	.29		.32	.88	.93	.98
.01	.95	-.12		.12	.89	.97	.28		.46	.85	.96	.98
.05	.95	-.01		.05	.93	.98	.30		-.08	.84	.96	.98
.10	.95	-.73		.21	.90	.98	.30		-.08	.79	.95	.99
.25	.96	-.62		.40	.95	.98	.50		.09	.84	.92	.99
.30	.97	-.62			.96	.98	.49		.34	.73	.83	.99
.35	.98	-.99			.94	.99	.49		.34	.64	.77	.99
.40	.99				.94	.99	.41		.08	.18	.62	.99
Number of Items												
All Items	60	25		25	30	50	25		25	30	60	40
.01	54	19		19	25	49	20		20	23	50	26
.05	39	15		16	16	33	15		13	18	39	22
.10	26	11		10	12	27	14		12	14	26	16
.25	17	6		4	7	14	7		5	11	12	12
.30	13	6		2	6	13	6		4	8	10	9
.35	8	4		1	4	8	6		4	6	8	6
.40	5	1		1	4	5	4		3	4	6	6



TABLE 3  
Correlation of  $z$  Item Difficulty Estimates

Acceptable Item Difficulties	Comparison and Type of Analogies									
	Word		Picture		Number of Items		Symbol		X	
	I	II	III	IV	V	VI	VII	VIII	IX	X
All Items	.96	-.08	.04	.91	.97	.30	.33	.89	.97	.98
.20-.80	.91	-.15	-.38	.76	.92	.21	.15	.79	.90	.92
.30-.70	.76	.20	-.12	.24	.89	.41	-.40	.60	.85	.83
.40-.60		-.50	-.16		.60					
All Items	60	25	25	30	50	25	25	30	60	40
.20-.80	39	18	22	16	31	17	20	17	25	20
.30-.70	19	8	19	6	20	9	7	10	17	13
.40-.60	2	4	7	1	5	1	1	2	3	3

vice employees on 25 picture analogies) showed an initial increase in invariance when those items with Rasch probabilities below .01 were removed, but when those below .05 were removed, the correlation fell to near-zero, and fluctuated randomly with subsequent deletions of items. In comparison IX (between college and high school students on 60 number analogies) the correlation increased when items with Rasch probabilities below .01 were deleted, and remained stable until after deletion of items with Rasch probabilities below .25. At that point the correlation began to drop.

The rest of the comparisons showed some increase in variance as items with low Rasch probabilities were deleted. In comparison IV (between college students and civil service employees on 30-word analogies), the increase in correlation was somewhat erratic. In comparison II (between college students and DVR clients on 25-word analogies), the item easiness estimates were negatively correlated. But this latter comparison and the comparisons of college and high school students on 60-word analogies (comparison I), on 50-picture analogies (comparison V), and on 40-symbol (comparison X), all correlated .99 when items with low Rasch probabilities were removed.

The relationship was relatively more simple for the  $z$  item difficulty estimates. In general, the less restrictive the range of acceptable item difficulties, the higher the correlations. For each of the six comparisons in which the  $z$  item difficulty correlated .90 or higher (comparisons I, IV, V, VIII, IX, and X), the highest correlations were observed when all items were included in the comparison, and the correlation dropped with each restriction of the range of acceptable item difficulties. The correlations fluctuated randomly with each restriction of the range of acceptable item difficulties for the four remaining comparisons (II, III, VI, and VII).

### *Test Calibration*

In estimating the amount of ability indicated by raw scores on a test, it is claimed that the Rasch model takes account only of the easiness of the items in a test. It is appropriate, therefore, to ask whether the ability estimates are invariant with respect to the ability of the calibrating sample. For each of the ten comparisons investigated (see Table 1), the product-moment correlation between the Rasch ability estimates was .999. Figure 1 illustrates the relationship between the ability estimates calculated for a 25-item word analogy test from the responses of 630 college students and the responses of 90 DVR clients (comparison II).

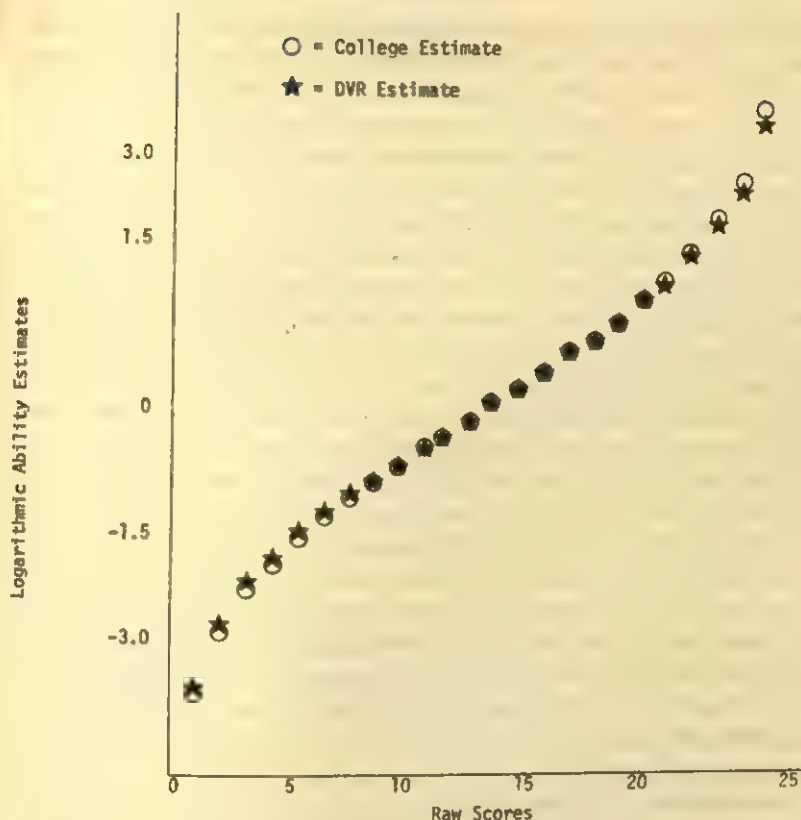


Figure 1. Invariance of Rasch Ability Estimate.

### Discussion

#### Item Calibration

Ten tests of the invariance of Rasch item easiness estimates were made with mixed results. However, the results are not as equivocal as they may appear. Anderson et al. (1968) have pointed out that the Rasch model does not lend itself to small samples. Generally samples of 500 or larger are needed to obtain stable item easiness (and ability) estimates. It is important, therefore, to keep the size of the sample in mind when interpreting the results. Comparisons I and X were based on the responses of 630 college students and 300 high school students and yielded correlations of .95 and .98. Correlations of .97 and .93 were obtained when the results obtained from 492 college students

were compared with those obtained from 120 and 145 high school students (comparisons V and IX). And comparisons IV and VIII based on the responses of 276 college students and 269 civil service employees yielded correlations of .91 and .88. In contrast, comparisons II and III involving item easiness estimates obtained from 89 DVR clients resulted in zero correlations.

Two comparisons (VI and VII) remain, however, which did not support the hypothesis under test. Both were based on small samples, but the samples were larger than some used in comparisons which did support the hypothesis. It is possible that the nature of the test was a factor in these results. Both comparisons involved the 25-item picture analogy test. It seems likely, therefore, that some factor other than ability and item difficulty may have influenced the probability of a correct response. This factor might have been recognition of some of the picture analogies as identical to the preceding word analogies.

Another factor which may have served to reduce the invariance of the item easiness ratios must be mentioned briefly. Panchapakesan (1969) provided a criterion for the elimination of examinees with low scores so that the estimation of item easiness will not be contaminated by guessing. According to her criterion some of the subjects in this study should have been eliminated. Because of the initially small sample sizes, this procedure was not followed. It is possible, therefore, that guessing may have reduced the invariance of the item easiness ratios in some instances.

In summary, six of the ten comparisons supported the hypothesis that the Rasch item easiness ratios were invariant with respect to the ability of the calibrating sample even though a number of the comparisons involved samples of questionable size. Of the four remaining comparisons, two included samples so small as to invalidate the results while the other two may have been invalid because the Rasch model was not appropriate for tests designed in that manner.

It must be noted that the results for the  $z$  item difficulty ratios followed those for the Rasch item easiness estimates. The data in the present study provide no basis for choosing between the two item parameters. Such a choice could be made, however, on the basis of the assumptions involved in the use of the two parameters. The  $z$  item difficulty estimate requires the assumption that the ability be normally distributed while the Rasch item easiness estimate requires no assumption about the ability of the calibrating sample. Therefore, either the samples used in this study were normally distributed in terms of ability, or  $z$  item difficulty estimates are robust for the assumption of normality.

The above results represent a stringent test of the Rasch model in

that items for which the Rasch model is clearly inappropriate were included in the comparison. Deletion of these items should result in an increase in the correlation of the item easiness estimates obtained from different samples. This result was observed for five of the six valid comparisons. In three of these comparisons (I, V, and X), the correlation increased to .99. In the other two cases (comparisons IV and IX), the correlation increased at first and then decreased. In both such instances the number of items remaining had become so small that the decrease in correlation may have resulted from a restriction of the range of item easiness estimates. Only comparison VIII (between civil service employees and college students on 30 picture analogies) failed to support this hypothesis. Both samples completed these picture items after completing 30 word analogies having identical relationships, thereby possibly contaminating their response to the picture analogies.

### *Test Calibration*

The results of each of the ten comparisons supported the hypothesis that Rasch ability estimates are invariant with respect to the ability of the calibrating sample. Even in those instances in which the samples were so small that the individual item easiness estimates were sample dependent, the resulting ability estimates were invariant. These results indicate that the ability estimates assigned to any collection of 25 or more items will be invariant with respect to the ability of the calibrating sample regardless of whether the separate item easiness estimates are invariant.

### *Conclusions*

The results of this research support the following conclusions:

1. Rasch item easiness ratios are invariant with respect to the ability of the calibrating sample when a sample of adequate size is used.
2. Invariance of the Rasch item easiness ratios is related to the goodness-of-fit of the items to the Rasch model. The deletion of items with low Rasch probabilities increases the invariance of the Rasch item easiness ratios.
3. The estimation of the amount of ability indicated by the raw scores on a test is invariant with respect to the ability of the calibrating sample for tests of 25 or more items, even when relatively small samples are studied.



## REFERENCES

- Anderson, J., Kearney, G. E., and Everett, A. V. An evaluation of Rasch's structural model for test items. *The British Journal of Mathematical and Statistical Psychology*, 1968, 21, 231-238.
- Barth, W. H., Lele, K., and Rosse, R. *Item analysis by the Rasch model*. Minneapolis: Department of Psychological Foundations of Education, 1970.
- Birnbaum, A. Efficient design and use of tests of mental ability for various decision making problems. School of Aviation Medicine, United States Air Force, Report No. 58-16, 1957.
- Birnbaum, A. Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology*, 1969, 6, 258-276.
- Bliss, E. F. The difficulty of an item. *Journal of Educational Psychology*, 1929, 20, 63-66.
- Brink, N. E. Overview of the Rasch model. *Proceedings of the 79th Annual Convention of the American Psychological Association*, 1971, 101-102.
- Brooks, R. D. *An empirical investigation of the Rasch ratio-scale model for item difficulty indexes*. (Doctoral dissertation, University of Iowa) Ann Arbor, Michigan: University Microfilms, 1965, No. 65-434.
- Guilford, J. P. *Psychometric methods* (2nd ed.). New York: McGraw-Hill, 1954.
- Guilford, J. P. Three faces of intellect. *American Psychologist*, 1959, 14, 469-479.
- Guilford, J. P. Intelligence: 1965 model. *American Psychologist*, 1966, 21, 20-26.
- Gulliksen, H. *Theory of mental tests*. New York: John Wiley and Sons, 1950.
- Helmstadter, G. C. *Principles of psychological measurement*. New York: Appleton-Century-Crofts, 1964.
- Horst, A. P. The difficulty of a multiple choice item. *Journal of Educational Psychology*, 1933, 24, 229-232.
- Lord, F. M. and Novick, M. R. *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968.
- Panchapakesan, N. The simple logistic model and mental measurement. Unpublished Doctoral dissertation, University of Chicago, 1969.
- Rasch, G. *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research, 1960.
- Rasch, G. On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Symposium on Mathematical Statistics*, 1961, 4, 321-334.
- Rasch, G. An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, 1966a, 19, 49-57.
- Rasch, G. An individualistic approach to item analysis. In P. F. Lazarsfeld and N. W. Henry (Eds.), *Readings in mathematical*

- social science*. Chicago: Science Research Associates, 1966b, 89-108.
- Richardson, M. W. The relationship between the difficulty and the differential validity of a test. *Psychometrika*, 1936, 2, 33-49.
- Schmidt, W. H. Necessity of the model. Paper given at the 1970 American Educational Research Association Presession on Sample Free Item Analysis and Person Measurement, Minneapolis, March, 1970.
- Sitgreaves, R. Review of G. Rasch, *Probabilistic models for some intelligence and attainment tests*. *Psychometrika*, 1963, 28, 219-220.
- Thorndike, E. L., Bergman, E. O., Cobb, M. V., and Woodyard, E. *The measurement of intelligence*. New York: Bureau of Public Teachers College, Columbia University, 1926.
- Thurstone, L. L. A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 1925, 16, 433-451.
- Thurstone, L. L. The calibration of test items. *American Psychologist*, 1947, 2, 103-104.
- Tinsley, H. E. A. *An investigation of the Rasch simple logistic model for tests of intelligence or attainment*. (Doctoral dissertation, University of Minnesota), Ann Arbor, Michigan: University Microfilms, 1972, No. 72-14, 387.
- Tucker, L. R. Selecting appropriate scales for tests. *Proceedings of the 1952 Invitational Conference on Testing Problems*. Princeton, N. J.: Educational Testing Service, 1953, 22-28.
- Wright, B. Sample-free test calibration and person measurement. *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton, N. J.: Educational Testing Service, 1967, 85-101.
- Wright, B. and Panchapakesan, N. A procedure for sample-free item analysis. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1969, 29, 23-48.
- Wright, B. and Panchapakesan, N. *Item analysis by the Rasch model, UCSL801*. Chicago: University of Chicago Computation Center, Social Science Program Library, 1970.



## ATTEMPT TO CONSTRUCT A SCALE FOR THE MEASUREMENT OF THE EFFECT OF SUGGESTION ON PERCEPTION<sup>1</sup>

V. A. GHEORGHIU, V. HODAPP, AND C. M. LUDWIG

University of Mainz  
Academy for Social and Political Sciences, Bucharest

A scale based on experimental methods has been prepared for measuring the effects of indirect suggestion upon perception. Three categories are included: (1) distorting the interpretation of presented stimuli, (2) inducing sense-impressions in the absence of adequate stimuli, and (3) producing insensitivity to stimuli that are objectively present. Test situations were designed for tactual, auditory, and visual perception. The scale was tested on a sample of 112 students from the 11th and 12th grades of a large city high school (58 girls and 54 boys).

Most of the item intercorrelations were positive and many significantly so. Eliminating the 9 lowest items of 21 left 12 for a reduced matrix, with the first factor accounting for 23% of the total variance. There were no special factors attributable to sensory modality. By summing the scores on the 12 items, a scale was produced with a reliability of .82.

Difficulties and limitations are discussed, along with the potential applications of such a scale in the study of socially important behaviors.

In his review of suggestibility in the normal waking state, Evans

<sup>1</sup> From the Department of Psychology of the University of Mainz, in cooperation with the Department of Psychology of the Academy of Social and Political Science, Bucharest. The addresses of the authors are as follows: V. A. Gheorghiu, Institutul de Psihologie, Str. Frumoasa 26, Bucuresti, Romania; V. Hodapp, Institut für Medizinische Statistik und Dokumentation der Universität Mainz, 65 Mainz, Langenbeckstr. 1, West Germany; C. M. Ludwig, Psychologisches Institut der Universität Mainz, 65 Mainz, Saarstr. 21, West Germany.

The sojourn of the senior author in Mainz during the course of the investigation was made possible through the German Research Society (Grant We 2/25, Fr 132/9).

We wish to acknowledge our gratitude to Professor Ernest R. Hilgard for assisting in the revision and preparation of the final draft of the manuscript.

Copyright © 1975 by Frederic Kuder

(1967) found that the reported experiments had explained very little of the basic phenomena. We agree with his assessment of the situation (see also the related discussion by Gheorghiu, 1972, 1973). For some time there has been considerable overlap in the investigations of suggestibility phenomena, particularly in sensory and motor tests (e.g., body-sway, arm and hand levitation, heat-illusion, progressive lines). Although these tests have been taken as a basis for understanding suggestion, they have in fact served merely to define the range of the behavior to be investigated. The paucity of firm conclusions is evidenced also in the factor analytic studies. As the investigations of Hammer and others (1963) and of Evans (1967) have shown, variations in the tests produce different factors of suggestibility.

The correlational and factor-analytic studies lead to the questions, first, whether the results show any clear unity or coherence within suggestibility phenomena, and, second, whether the measures meet appropriate experimental standards relevant to the behavior under investigation. Heavy weight has been given in recent years to the interpretation of statistically detected factors, and little attention has been paid to developing and adapting behavioral measures based on underlying hypotheses about suggestibility and its manifestations.

To the best of our knowledge, despite all the investigations that have been made both clinically and experimentally, there is not a single standardized test battery to be found. The clarification that can be expected from a careful application of experimental methods is lacking.

Looking back on the findings of different authors, gross differences are found, leading Duke (1964) to say that it is essential "that all details of the test situation be carefully examined, specified, and appraised." This is without doubt necessary, but it seems equally important to be clear about the definition of what is measured by the suggestibility tests, and to specify the purposes intended by their investigation. Our primary purpose has been to investigate interpersonal differences in the effects of suggestion on perception. With this background it will be possible in later investigations to relate suggestion to other influences upon perception, and to consider the findings in the framework of differential psychology.

Although a complete sensory suggestibility test battery would have some variants in the kinds of suggestions used (direct, indirect, co-judge, etc.) we have chosen to limit our first scale to perception using the single variant of indirect suggestion. There are three main categories of the influence of suggestion upon perceptions: (1) distorting the interpretation of presented stimuli; (2) inducing sense-impressions in the absence of adequate stimuli; and (3) producing insensitivity to stimuli that are objectively present. Many of the current



investigations of suggestibility have not carefully distinguished between these functions, and have not studied the functions systematically in reference to various sensory modalities.

With the proposed scale, in which the functions studied are carefully specified and experimentally measured, further experiments can be performed to investigate many aspects of influences on cognitive processes such as imagination, expectation-deception, agreement tendencies, etc. Preliminary experiments, previously reported (Gheorghiu, Hodapp, and Thiedig, 1972) suffice to show that it is possible to prepare such a scale.

### *Experimental Design*

In order to carry out the plan for developing a scale that would test the effect of suggestion on perception, experimental test situations were designed for tactual, auditory, and visual perception, producing as far as possible analogous demands upon the subject, and having in mind the three categories previously mentioned. In all, seven tests were designed for each modality. The 21 tests are described in detail in Table 1.

To make some of the sensory distortions more plausible, all tests were conducted with sensory input attenuated or impeded. We introduced an appropriate "impediment" for each sense: plaster on the skin for the area to be tactually stimulated, a foam-rubber insert in the headphones, and dark glasses before the eyes. The main purpose in using the impediments was to increase the subject's concentration of attention, and to make the suggested sensory experiences more credible.

### *Subjects*

Pupils from the 11th and 12th grade of a large city high school served as subjects in the investigation. With few exceptions, entire classes were tested during school hours, although each test was given individually. The total sample of 112 consisted of 58 girls and 54 boys. Ages ranged between 15 and 18 years. The majors were as follows: science, 38%; language arts, 47%; and social studies, 15%.

### *Instructions*

All instructions were given by tape recorder. The investigation was explained to the subject as a study of perception thresholds, accommodation, and discrimination of stimuli. It was explained that the subject might perceive something or perhaps nothing. His main task

TABLE 1  
*Experimental Designs (Items 1-3)*

	Tactile	Auditory	Visual
Item 1: simulation of an increasing stimulus	The subject puts one hand through the aperture of a partition wall. A beaker containing sand is put in the hand. It is pretended that more sand is put into this beaker.	A tone, produced by a tone-generator, is presented to the subject through one side of a pair of headphones. This tone is seemingly slowly increased by manipulation of the device.	The light intensity of a bulb is seemingly increased by manipulation of joined transformer.
Item 2: giving a meaning to an unspecific stimulus	Behind the partition wall an abstract plastic wire shape is pressed on the subject's hand. The subject has to identify it as a number between zero and five.	Indistinct background noise recorded on tape is presented to the subject through some headphones. The subject has to distinguish a spoken number between zero and five from this noise, although no number was actually spoken.	One of the cards of the ISHIHARA-test is presented to the subject as a photocopy (black and white), so that actually no number can be perceived within it. The subject should discover a number between zero and five from this card.
Item 3: simulation of a connection between two different stimuli of which only one is given	A connection between the light intensity of a bulb and the electromagnetically produced vibration of a string is demonstrated to the subject. In the experiment, however, the string remains motionless, while the bulb becomes more and more bright.	The vibration tone of a string is presented on the background of a white (pure?) noise through the subject's headphones. In the experiment the string is moved, but the tone is not sent through the headphones.	It is demonstrated to the subject, how a bulb becomes brighter as the turning frequency of a joined dynamo becomes greater (i.e. the tone of the dynamo becomes louder) which seemingly leads electric current to the bulb. In the experiment, however, the bulb is switched off.

	Tactile	Auditory	Visual
Item 4: simulation of a sensory stimulus without previous objective stimulation	The subject puts one hand into a box through which a pin is let down to the hand by a visible cord. In the experiment, however, the pin is drawn to the back by a special catch mechanism inside the box, so that the pin cannot reach the hand of the subject.	A stop watch on a stand is brought near to the subject's ear. The subject has to say, when he can hear the watch ticking, which however, is switched off in the experiment.	On a sideways a black cardboard disc is brought near to the eye of the subject. The subject is told that there is a red point in the middle of the disc.
Item 5: simulation of a sensory stimulus with previous objective stimulation	The subject puts his hand into the box. Three rings of decreasing weights are let down by the cord, but only the first two rings reach the hand of the subject.	Three tones of decreasing intensities are said to be presented to the subject by an audiometer. An objective stimulation is actually effected only in the first two cases.	On the sideways three black cardboard discs are brought near to the eye of the subject. In the middle of them there are colour points of decreasing sizes and light intensity however, in the experiment, the back of the third disc which is totally black is presented the third time.
Item 6: simulation of a bilateral stimulus with an only one-sided given stimulus	Behind the partition wall both hands of the subject are said to be touched with pins at the same time. Actually only one pin reaches the hand of the subject.	At the same time two stop watches are brought near to the ears of the subject. Only one watch is switched on.	Behind a device, which separates the fields of vision of the subject two light sources are brought near on the sideways. Only one light is actually switched on.
Item 7: simulation of a stimulus annulment	Behind the partition wall some water is taken out of a water-filled beaker which stands on the subject's hand by a syringe. The subject has to say, when he cannot feel the beaker on his hand any more.	Ticking of a switched on stop watch recorded on tape is presented to the subject through the headphones. The intensity is said to be decreased by manipulation of the volume control of the tape recorder. The subject has to say, when he does not hear any more ticking.	The light intensity of a bulb is said to be reduced by manipulation on a seemingly joined transformer. The subject should say, when he cannot perceive the light stimulus any more.

would be to concentrate with utmost attention on what he perceived, if, and with what degree of certainty, he experienced the sensation.

### *Procedures*

The functioning of each of the devices was demonstrated before every test, without, however, stimulating the subject directly. Each single task began with the signal "Attention" and lasted 15 seconds. The subject was instructed to say "Now" as soon as he perceived the stimulus. The reaction time between presentation and his saying "Now" was recorded. The subject was then asked to state whether his perception was "certain" or "uncertain." The items were always presented in the same order within a sensory modality, but the modalities were presented in random order. Each item was presented twice, once to one hand, ear, or eye, next to the other hand, ear, or eye. Total testing time was between 40 and 50 minutes.

There were two male and two female experimenters (psychology students) who worked independently.<sup>2</sup> Each experimenter tested an equal number of male and female pupils, the same experimenter conducting all tests with the same pupil.

### *Results*

#### *Item Reliability on Retesting*

As indicated, we carried out two trials for each item. The computed Phi-coefficients based on the simple dichotomy for each item of "reaction/no reaction" lie between 0.25 and 0.57 with two exceptions only (Tactile Item 4 and Tactile Item 5). All the coefficients are positive, and can be considered as reliability coefficients. When only one of the two items was passed there appeared to be no tendency for the first or the second to be more frequently passed, as determined by the  $X^2$  (chi-square) test of McNemar.

#### *Item Intercorrelations and Factor Analysis*

If there is some common or unified aspect of suggestibility running through the tests, it is to be expected that the scores will be positively correlated. For each item the score could be 0, 1, or 2, depending on whether the response failed to occur or occurred on one or both trials. When all of the intercorrelations were examined, it was found that

<sup>2</sup> The help of E. Feingold, K. Krein, G. Ries, and Chr. Wortmann, who worked with us in conducting the experiments, is gratefully acknowledged.

most of them were positive, many of them significant, and of the few that were negative, none was significant. In order to emphasize the common factor, 9 of the 21 items yielding the lowest correlations with the total test were eliminated in the further analyses. The reduced matrix, based on the remaining 12 items, is given in Table 2. Most of the correlations are significant, and none is negative. It may be inferred from such a correlation matrix that there will be a common factor present. It was found that the latent root of the first principal component came to 4.74, representing 22.6% of the total variance, and further roots were negligible (Figure 1).

Multi-factorial solutions, rotated according to Kaiser's varimax criterion, failed to yield any factors unique to either the special sense modalities or the categories of suggestion. There are some technical problems in interpreting this finding, to which reference will be made later.

### *Item Analysis*

If all items are summed, as though there is a 21-item scale, the total scale yields a reliability coefficient of .80, according to the method Hoyt and Stunkard (1956), equivalent to the Kuder-Richardson alpha for weighted items.<sup>3</sup> Because a few of the items had correlations with the total test (part-whole corrected) well below .30, they were eliminated as noted previously, and 12 items retained. This 12-item scale had a reliability of  $\alpha = 0.82$ , according to Hoyt-Stunkard.

Table 3 shows the results of the item analysis for the selected items. Of the items retained, three (Items 1, 5, and 6) were common to all three modalities, one was common in addition to audition and vision (Item 4), and another (Item 3) was found for vision only. Items 2 and 7 were unsuccessful by this criterion for all sense modalities.

Although various weighting methods were tried out, the preferred method proved to be to use the sum scores as previously described, 0, 1, or 2 for each item. The distribution of the sum scores based on the 12 items was skewed to the right with a mean of 8.78 out of a possible 24. The standard deviation was 5.31, and the standard error of measurement 2.25. Hence there is a 95% probability that the "true test value" will fall within the limits of  $X \pm 4.41$ .

### *Subjective Certainty and Reaction Time*

When subjects were classified as more highly suggestible (with sum scores of 8 to 24) or less highly suggestible (with sum scores of 0 to 7),

<sup>3</sup> For the statistical methods employed, see Lienert (1969). The item analyses were conducted by means of Program IT 09, H. Vorkauf, Deutsches Rechenzentrum Darmstadt.



TABLE 2  
Correlation Coefficients

	1T	5T	6T	1A	4A	5A	6A	1V	3V	4V	5V	6V
1T												
5T	0.13											
6T	0.23*	0.23*										
1A	0.31**	0.17	0.32**									
4A	0.17	0.18	0.24*	0.35**								
5A	0.08	0.30**	0.12	0.25**	0.29**							
6A	0.26**	0.20*	0.33**	0.32**	0.43**	0.31**						
1V	0.18	0.07	0.28**	0.34**	0.24*	0.12	0.33**					
3V	0.10	0.23*	0.46**	0.35**	0.21*	0.37**	0.40**	0.36**				
4V	0.14	0.20*	0.20*	0.27**	0.21*	0.19	0.32**	0.30**	0.38**			
5V	0.07	0.15	0.27**	0.19	0.38**	0.28**	0.32**	0.31**	0.43**	0.47**		
6V	0.17	0.20*	0.40**	0.32**	0.35**	0.34**	0.41**	0.26**	0.57**	0.46**	0.38**	

\*  $p < 0.05$ .

\*\*  $p < 0.01$ .

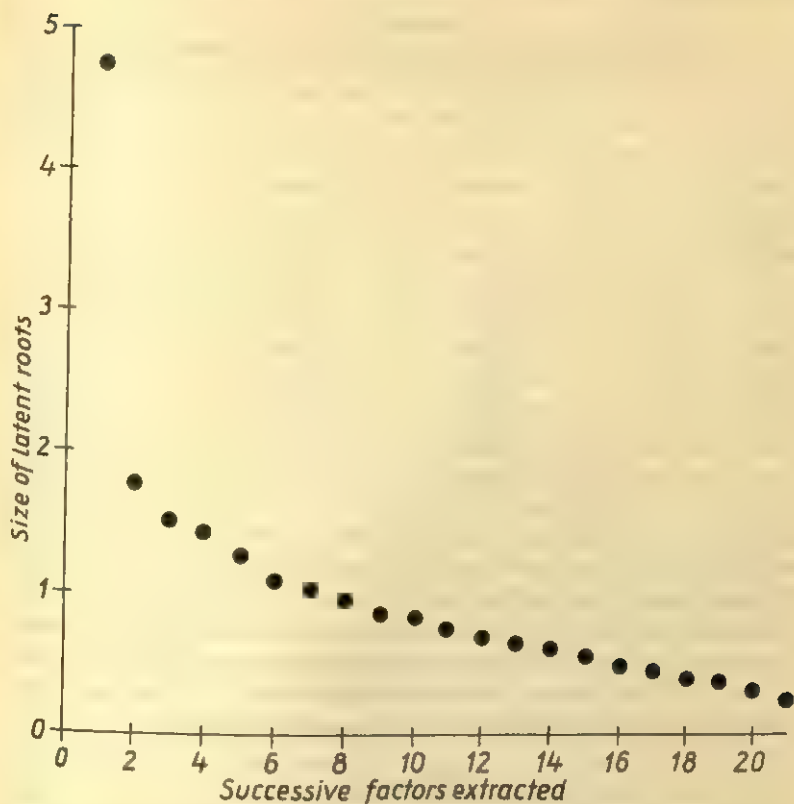


Figure 1. Latent roots of the principal components extracted from the matrix of correlations of the 12 items (Table 2).

"certain"-reactions were significantly more frequent among the highly suggestible, but they predominated for both groups.

In general, those subjects who reacted "uncertain" took relatively more time to respond than those who reacted "certain," but the relationship between "certainty" and speed of reaction was not strong, and did not show up as a connection between the sum scores and reaction time.

#### *Experimenter Effect and Sex of Experimenter and Subject*

The four experimenters produced somewhat different results. The raw scores for each subject were transformed by the  $\log(1 - X)$  to eliminate the dependence between mean values and the variances of the groups of subjects (Table 4). A simple analysis of variance of these transformed scores, classified by experimenter, resulted in an  $F$ -value of

TABLE 3  
*Characteristic Values of Item Analyses of the 12 Selected Items*

Item	mean value	stand. deviation	item-test correlation
1 Tactile	1.49	0.72	0.28
5 T	0.35	0.56	0.31
6 T	0.84	0.84	0.48
1 Auditory	0.98	0.79	0.50
4 A	0.74	0.82	0.48
5 A	0.62	0.75	0.40
6 A	0.62	0.79	0.58
1 Visual	1.08	0.85	0.44
3 V	0.54	0.76	0.61
4 V	0.36	0.66	0.49
5 V	0.70	0.85	0.51
6 V	0.46	0.70	0.61

2.84 ( $f_1 = 3, f_2 = 108, p < .05$ ). Although the female experimenters (3 and 4) produced the higher mean scores, the difference of scores by sex of experimenter was not significant (according to the Scheffé method).

The point-biserial correlation between sex of the subject and the raw sum score was .14, indicating essentially that the scale is independent of the sex of the subject. The interaction between the sex of the subject and the sex of the experimenter also proved nonsignificant.

### Discussion

The main result of this investigation was that a consistent scale could be constructed to measure suggestible behavior in the area of perception, using an indirect approach.

Although the factor analysis yielded a first factor accounting for 23% of the variance, and multiple-factor methods led to no meaningful subordinate factors, this indication of a unity or coherence underlying the measures must not be accepted uncritically because of the manner

TABLE 4  
*Mean Values and Variances of the Sumscores, Separated According to Experimenters (N = 28 for Each Experimenter)*

Experimenter	mean value	variance
male		
1	8.07	26.37
2	6.86	13.76
female		
3	8.64	21.79
4	11.54	41.74

in which test items enter into a determination of factors. It appears quite clear that there are no special factors attributable to sensory modality, because the three modalities were amply represented in the tests. However, the three types of influence upon perception were not equally represented, and it appears that the common factor can be considered more representative of perceptual distortion than of either producing a perception in the absence of a stimulus, or of negating a stimulus objectively present. Items 1, 5, and 6, that appear in the final scale for all sensory modalities, represent, first, an increasing stimulus, second, the continuation of a series following objective stimulation, and third, the supplementation of a single stimulus by its bilateral representation. These appear to fit the distortion paradigm better than the pattern of creating a perception in the total absence of sensory stimulation. Because more items are of this kind than of pure representations of creating a perception or of completely annulling one, it may be that the appearance of special factors is in part owing to the relative frequency of acceptable items in the three categories. The one item clearly reflecting annulment of a stimulus (Item 7) does not appear in the final scale at all because it correlated too low with the total test. If, however, there had been several such low correlations correlated with each other, they would then have determined another factor. A clarification of this problem is a task for the future.

In the item analysis the visual items turned out to be more representative of the scale as a whole than the items of the other senses, with five visual items appearing in the 12-item scale as against four auditory and three tactile items. The advantage for the visual items may owe to the better control of conditions of testing, including darkening of the experimental room as a possible support for the influence of suggestion.

The influence of the experimenter upon the results is somewhat disturbing. With a small sample of experimenters the result could well have been due to slight differences in the manner in which the tests were presented to the subjects. Therefore it seems necessary, especially in suggestibility research, to unify and control the experimental performances even more carefully than we did.

The experiment raises a number of questions that can be answered only by further investigation. For example, it is not known whether or not the "certain" and "uncertain" responses are influenced by the personality of the subject, and the findings on reaction time suggest that the optimal duration of the experimental item should be studied.

Because these experiments have dealt with only the indirect variant of suggestion, no definite statements can be made about the components that cause the given response. It is an open question whether we

are dealing primarily with the ability to imagine or, instead, with a tendency to agree or to comply. The possibility of social compliance is always present when one relies, as in these experiments, on the verbal reaction as the indicator of suggestible behavior. It will therefore be necessary to carry out other experiments with other variants to determine which components are operative.

The present scale, supplemented by other variants, should ultimately provide a measure of individual differences in known behavior that can then be used in studying other socially important behaviors, whether imagination, tendencies to agree or disagree, or responsiveness to expectations of various kinds.

### Summary

1. The experiments yielded a scale that measured the influence of suggestibility upon perception consistently in the areas of tactile, auditory, and visual perception.

2. By weighting the responses according to whether one or both of the tests was passed with repeated items, a 12-item scale with an internal consistency of 0.82 resulted.

3. The degree of reaction was also differentiated by the subject's report of "certain/uncertain" following each perceptual report.

4. The data indicated a significant experimenter effect. Although the two female experimenters produced numerically higher scores than two male experimenters, the difference was not significant. We also found no significant differences based on the sex of the subjects.

### REFERENCES

- Duke, J. D. Intercorrelational status of suggestibility tests and hypnotizability. *Psychological Record*, 1964, 14, 71-80.
- Evans, F. J. Suggestibility in normal waking state. *Psychological Bulletin*, 1967, 67, No. 2.
- Gheorghiu, V. A. Betrachtungen über Suggestion und Suggestibilität. (On suggestion and suggestibility). *Scientia*, 1972, 107, 811-860.
- Gheorghiu, V. A. *Untersuchungen zur sensorischen und motorischen Suggestibilität*, unveröffentlichte Habilitationsschrift, Mainz, 1973.
- Gheorghiu, V. A., Hodapp, V., and Thiedig, S. Untersuchungen zur taktilen, auditiven und visuellen Suggestibilität. *Archiv für Psychologie*, 1972, 124, 303-320.
- Hammer, A. G., Evans, F. J., and Bartlett, M. Factors in hypnosis and suggestion. *Journal of Abnormal and Social Psychology*, 1963, 67, 15-23.
- Hoyt, C. I. and Stunkard, C. L. Estimation of test reliability for unrestricted item scoring methods. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1956, 12, 756-758.
- Lienert, G. A. *Testaufbau und Testanalyse*. Weinheim, 1969.



## A STUDY OF THE EFFECT OF THE VIOLATION OF THE ASSUMPTION OF INDEPENDENT SAMPLING UPON THE TYPE I ERROR RATE OF THE TWO-GROUP $t$ -TEST

ROBERT W. LISSITZ AND STEVE CHARDOS

University of Georgia

This paper describes some of the situations in which a psychologist is likely to violate the assumption of independent errors. A Monte-Carlo study of the effects of this violation is then described. A number of examples of different kinds and degrees of dependency are included along with a table that gives the effect of the dependency upon the shape of the test statistic's distribution. The results of this study demonstrate that this assumption is a critical one. Researchers are strongly urged to avoid hypothesis testing if they suspect that the assumption of independence has been violated.

THIS paper is an attempt to examine the effect upon the type I error rate of the violation of the assumption of independent random sampling within the specific context of the two group  $t$ -test. It is our feeling that the results generalize far beyond this situation, but for concreteness we will consider this case in some detail. This is an assumption that has been largely ignored in the statistical literature concerned with robustness. Boneau (1960) has written what is probably the most popular article for psychologists on the subject of assumptions of the  $t$ -test. His article does not treat the independence problem at all. There are probably some reasons for the lack of interest in the subject of this assumption but it is our feeling that this neglect is most unfortunate.

One reason that we feel that the assumption needs study is that it is a common problem in psychological applications. For example, consider a study in which the experimenter is interested in the behavior of subjects who are participating in therapy groups. He might run three groups under one condition and three groups under another condition

and then treat the design to a *t*-test. Instead of there being three experimental units (one for each group) he uses the individual scores from each person in the group. These individual scores are, of course, dependent upon one another. Another example of a situation in which dependency can arise is one in which the subjects are volunteers (such as an introductory psychology class). If the subject likes the experiment the subject will tell his friends and they will come to the experiment with a certain set that is obtained from the first subject. This introduces a certain level of dependency among subjects. Another example is one in which the dependent variable results from a rating procedure in which a single rater looks at more than one of the subjects, again introducing a dependency.

Another reason for the importance of studying this assumption is that it appears in most statistical tests. A common criticism of metric statistics is that they make assumptions that are untenable and then the critic suggests using nonparametric or distribution free statistics. Unfortunately these alternatives also make the assumption of independence across subjects.

The effect of making a variety of assumptions has been the subject of two excellent, recent reviews. One of these reviews is by Glass, Peckham, and Sanders (1972) and the other by Huber (1972). As noted by Glass, et al. (1972), there have been two papers and a book that are concerned with the problem of dependency. G.E.P. Box (1954) has discussed some of the problems related to analysis of variance and the effect of violation of assumptions. Sometimes the violation of assumptions cannot be avoided by more careful experimental design. As he states (page 484) "Data occur, however, in circumstances where there is no possibility of using this device (randomization), usually because the factor which is to be studied is the effect of time or position, which itself gives rise to the correlation." He considers the situation in which there is a serial correlation between errors and provides a table which indicates that the true type I error will far exceed the nominal value provided by the experimenter when there is even a relatively small positive correlation. Cochran (1947) indicates that a constant correlation between every pair of scores can have a large effect upon the true variance of the treatment mean and, therefore, an effect upon the resulting test statistic. He does not pursue this problem in any great detail and does not present data in terms of the type I error rate.

The book by Scheffé (1959) has a short section on the effect of violation of the assumption of independence of errors. Scheffé also deals with the case of serial correlation and briefly describes the effect upon confidence intervals of the mean under large sample conditions. He presents a table of the true alpha error rate that indicates that with

positive serial correlation the alpha rate goes up and with negative serial correlation it goes down.

These few references are all that are mentioned in the two recent reviews referred to above. There is additional work but it is preliminary and/or impossible to apply to most psychology problems, at the present time.

The following sections of this paper describe a Monte-Carlo study in which the effect of nonindependence is analyzed. It is hoped that the following series of realistic examples will illustrate the need for more careful attention to this assumption.

### *Procedure*

A FORTRAN IV (double precision for the CDC 6400) computer program<sup>1</sup> was written that would generate successive samples of data that are used to calculate two sample *t*-tests. A multivariate normal generator<sup>2</sup> was used that was extremely accurate even for a very large number of dimensions. The use of this generator allowed for the specification (and hence variation) of the covariance structure of the subjects within a group while maintaining the normality and equal variance assumptions as well as the null hypothesis of no difference in the means.

A mean vector of zeros (the number of elements equal to the number of subjects in a group) was input along with the variance-covariance structure desired.<sup>3</sup> The variance-covariance matrix always consisted of ones in the diagonal, thus making the off-diagonal values equivalent to correlations between subjects within a group, across replications.

A vector of scores was generated and each element was the score for a subject in the first group. A second vector was generated to give the scores for the second group. The standard equation for the *t*-test of the difference between two independent samples was then applied to these data and the resulting *t*-value tabled. This process was repeated 1000 times thus resulting in 1000 *t*-values that could be compared to the theoretical (expected) *t*-distribution.

<sup>1</sup> A copy of the complete program is available upon request from the first author at the Psychology Department, University of Georgia, Athens, Georgia, 30602.

<sup>2</sup> We would like to thank Dr. Rolf Bargmann of the Statistics Department, University of Georgia, for the generous loan of his excellent multivariate normal generator and for his help in operationalizing the subroutine.

<sup>3</sup> This means that the null hypothesis of equal means was true even though the covariance structure has been altered. In other words, the effect of the dependency among subjects is not systematic by group in such a way that one group will have a larger mean than the other. This is an important type of dependency, but the effect upon the probability of rejecting the null hypothesis is clear—it will increase this probability. In contrast, the effect of the types of dependency we are considering is not at all clear.

Because of the expense of computer time only one sample size was chosen for this illustration. The size selected was 31 subjects in each group giving equal sample sizes for the two groups and 60 degrees of freedom. This, of course, means that the multivariate normal generator had a mean vector with 31 elements (each zero) and a  $31 \times 31$  variance-covariance matrix. It was our feeling that this represented a "typical" application of the two group *t*-test.

### Results

The first set of *t*-values to be run consisted of the case in which the variance-covariance structure was an identity matrix. This was tabled as a check on the computer program. As can be seen in Table 1 the column titled "independent" resulted in a distribution of *t*-values that are very close to the expected set of *t*-values. The differences that did result are extremely small.<sup>4</sup>

A variety of types of dependency were investigated and are also a part of Table 1. The first are the cases in which every subject is correlated with every other subject either .2 (col. I), or .4 (col. II), or -.2 (col. III), or -.4 (col. IV): The variance-covariance matrix, in these cases, is a constant matrix except for the diagonals. These last two cases are nongramian and, therefore, impossible in practice, although the computer program was written to handle these cases. They are included here as idealizations of reality and because the results were sensible and very interesting. As can be seen, the effect of non-independence is extremely large even in the .2 case. The effect of positive dependence is to increase the size of the tails and the effect of negative dependency is to decrease the size of tails of the empirically derived distribution.

Another type of dependency that might be expected in psychological experiments is that of the serially correlated subjects. This set of variance-covariance matrices consists of zeros everywhere but the main diagonal, and the diagonal on either side. That is, adjacent subjects are correlated with each other and independent of all other subjects. Again, dependencies of .2 (col. V), .4 (col. VI), -.2 (col. VII), and -.4 (col. VIII) were run and tabled. This set of results agrees with those found by Box (1954) and Scheffé (1959) and the general conclusions are the same as for the earlier data except that the magnitude of the effect is less.

---

<sup>4</sup> The chi-square test of goodness of fit was barely significant at the .05 level, but (in our opinion), for the purposes of this paper, the results are quite clear and not distorted by the very minor divergence detected by the statistical test. It should be noted that the power of this test is very great since it involves 1,000 *t*-values.

TABLE I  
*Proportion of Monte-Carlo Generated t-Values in Each Interval of the t-Distribution*

Distribution of t values	Expected t Percent- age	Constant Covariance					Empirically Obtained Percentages					Decreases. 5.4.3.2.1 (XI)
		.2 (I)	.4 (II)	-.2 (III)	-.4 (IV)	.2 (V)	.4 (VI)	-.2 (VII)	-.4 (VIII)	.2 (IX)	.4 (X)	
$\infty$ to 2.390	1.00	20.60	31.20	0.00	0.00	2.20	6.80	.10	0.00	6.30	19.10	30.10
2.390 to 2.000	1.50	3.30	2.50	0.10	0.00	2.70	3.90	1.20	0.10	3.10	4.70	3.30
2.000 to 1.670	2.50	2.20	3.40	0.10	0.00	3.30	5.20	2.10	0.50	3.30	4.80	1.90
1.670 to 1.296	5.00	4.60	5.80	2.00	0.60	7.60	5.00	4.00	1.90	6.30	4.00	3.90
1.296 to .679	15.00	14.20	6.30	5.30	11.70	9.90	11.20	12.20	12.50	12.60	6.40	3.50
.679 to .254	15.00	14.70	6.10	3.70	20.80	23.70	11.10	16.90	18.50	10.40	4.80	3.90
.254 to 0.000	10.00	10.40	2.90	2.00	14.20	17.40	7.70	12.30	15.20	6.10	3.10	1.90
0.000 to -.254	10.00	7.90	2.20	1.70	16.70	18.50	5.60	11.50	14.00	5.90	3.40	1.40
-.254 to -.679	15.00	15.50	5.50	3.60	19.70	20.40	9.70	19.80	21.60	11.10	5.50	3.20
-.679 to -1.296	15.00	15.10	7.70	4.90	13.60	9.10	11.80	14.30	12.80	15.10	8.90	4.80
-1.296 to -1.670	5.00	5.10	4.50	2.70	1.10	0.40	9.00	3.30	2.00	6.60	5.40	2.90
-1.670 to -2.000	2.50	3.60	4.30	3.30	0.20	0.00	4.20	1.40	0.80	4.30	4.20	3.20
-2.000 to -2.390	1.50	2.20	4.90	3.20	0.00	0.00	3.40	.60	0.10	3.60	4.70	3.00
-2.390 to $\infty$	1.00	1.10	21.80	30.70	0.00	0.00	8.70	.30	0.00	5.30	21.00	33.00



A third type of dependency is a generalization of the serial type in which each subject is correlated with his two adjacent subjects. The covariance matrix has, in addition to the ones in the diagonal, constants of .2 (col. IX) in the adjacent two diagonals or .4 (col. X) in these adjacent two diagonals. Every other value is zero. The effect of this dependency is to increase the type one error rate. The amount of increase for the .2 case is roughly equivalent to the .4 serial case, and the .4 case is much larger.

A fourth type of dependency is that in which every subject is assumed to communicate with five other subjects and the amount of communication is further assumed to diminish across the five subjects. The covariance matrix was designed to reflect this. The values adjacent to the main diagonal are, from the left: .1, .2, .3, .4, .5, then the 1. in the diagonal and to the right, are: .5, .4, .3, .2, .1. This case, as can be seen in Table 1 (col. XI), gives rise to an increase in the true type I error rate that is comparable to the constant .4 dependency matrix.

### Discussion

The results section indicates quite clearly that the *t*-test is not robust to the assumption of independence. Even when as small as 4% of the total variance is shared between subjects (i.e., 96% independent) the effect upon the true alpha level is considerable. The appropriate conclusion for the user of this, and probably any other test, is to ignore the significance level if he has any reason to believe that there is a lack of independence.

Considerable literature on the general linear hypothesis involving a general variance-covariance matrix exists. Some of this work is presented in the book by Johnston (1972). A particularly important point for the user is that the linear model in which the disturbance terms are incorrectly assumed to have zero covariance does not lead to bias in the estimation of parameters of the model. Instead, the problems arise with the variance of the estimator. It is no longer minimum variance. The minimum variance unbiased estimator requires knowledge of the population covariance structure. The implication of this material seems to be that if the researcher is interested in just estimating parameters, he will be in less trouble than if he tries to use classical hypothesis testing.

### REFERENCES

- Boneau, C. A. The effects of violations of assumptions underlying the *t*-test. *Psychological Bulletin*, 1960, 57, 49-64.

- Box, G. E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems, II. Effects of inequality of variance and of correlation between errors in two-way classification. *Annals of Mathematical Statistics*, 1954, 25, 484-498.
- Cochran, W. G. Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics*, 1947, 3, 22-38.
- Glass, G. V, Peckham, P. D., and Sanders, J. R. Consequences of failure to meet assumptions underlying the analysis of variance and covariance. *Review of Educational Research*, 1972, 42, 237-288.
- Huber, P. J. Robust statistics: a review. *The Annals of Mathematical Statistics*, 1972, 43, 1041-1067.
- Johnston, J. *Econometric methods*. New York: McGraw-Hill, 1972.



## THE RELATIVE VALIDITY OF SCALES PREPARED BY NAIVE ITEM WRITERS AND THOSE BASED ON EMPIRICAL METHODS OF PERSONALITY SCALE CONSTRUCTION<sup>1</sup>

DOUGLAS N. JACKSON  
University of Western Ontario

In an effort to evaluate alternative strategies of personality scale construction, the extent to which relatively naive item writers could produce valid personality scales was investigated. Each of 22 undergraduate psychology students was asked individually to prepare 16 items for one of three scales: Social Participation, Tolerance, and Self-Esteem. These items were administered to a sample of 116 females, comprising pairs of roommates, together with like-named scales from the empirically-derived California Psychological Inventory (CPI) and from the Jackson Psychological Inventory (JPI). Self-ratings and roommate ratings on these trait dimensions which served as the two criterion measures were also obtained. Student scale validities, which were much higher than those obtained for the CPI, were almost comparable to those for the JPI. Student scale scores were less free from desirability variance than were JPI scale scores. Like the earlier Ashton-Goldberg study, the results were interpreted as supporting a construct-oriented over an external-empirical strategy of personality scale construction.

THE two major purposes of this study were (1) to appraise the extent to which relatively naive item writers can produce valid personality scales, and (2) to compare these scales with scales derived from other strategies, particularly those employing empirical methods using an

<sup>1</sup> Portions of this paper were presented at the meetings of the Canadian Psychological Association, Windsor, Ontario, June 12, 1974. Supported in part from a research grant from the Canada Council. Grateful acknowledgement is made to Cheryl Kuhwald, Patrick Buckley, Margaret Rintoul, and Esther Wagner for their assistance.

Copyright © 1975 by Frederic Kuder

external criterion. There has been a continuing concern (Hase and Goldberg, 1967; Goldberg, 1972) regarding the optimal method for constructing personality scales. Recently, the author (Jackson, 1971) issued a challenge to the effect that if the most elaborate empirical strategies of personality scale construction were pitted against the work of one or two item writers each of whom spent about two or three hours, validities would be higher for the scales constructed by the item writers. This point of view, by no means widely accepted, was issued as a challenge to researchers to conduct necessary comparative studies to determine whether and under what conditions the assertion might be accurate.

In an important study Ashton and Goldberg (1973) undertook to evaluate systematically the extent to which use of naive and novice item writers might be superior to other strategies of personality scale construction. They compared a set of scales prepared by two groups without extensive experience with personality scale construction—laymen and psychology graduate students—to sets of scales developed by the more elaborate empirical procedures drawn from the California Psychological Inventory (CPI). For the three scales investigated, Sociability, Achievement, and Dominance, peer-rating validities were consistently higher for personality scales developed by psychology graduate students than for those drawn from the widely-used California Psychological Inventory. Validities for scales developed by laymen, although significant in general, were considerably lower than those validities for scales prepared by the graduate students. Additionally, convergent validities of psychology students were found to be in a comparable range with the much more elaborately constructed Personality Research Form (Jackson, 1967), which employed a combination of rational and empirical procedures concerned with obtaining optimum levels of internal consistency and freedom from bias.

The present investigation, which was, in part, a replication and extension of the Ashton-Goldberg study, employed (a) a different set of personality constructs, the definitions of which are not quite so obvious to laymen as to professional item writers, (b) different methods for obtaining peer ratings, and (c) a somewhat different population of item writers. In addition to validity, personality scales were also evaluated for freedom from response style variance. It is hoped that by introducing procedural variations additional insight regarding the generalizability of the important Ashton-Goldberg findings will become possible. This procedure in turn should provide additional data bearing on the radically different approaches to scale construction proposed by Meehl (1945) and others and by Jackson (1971).



## Method

### Subjects

A total of 22 undergraduate psychology students, largely in their third year of university, comprised the set of item writers employed in the study. All were enrolled in the author's course in Psychological Tests and Measurement.

The respondents to the personality scales were 104 females drawn from a large women's residence at the University of Western Ontario, all of whom were paid volunteers. A prerequisite in the study was that each person have a roommate who was also willing to participate.

### Procedure

The student item writers were first introduced to the question of different methods of scale construction and to the hypothesis that novice item writers could produce personality scales with worthwhile validities. The three targeted personality traits, Social Participation, Tolerance, and Self-Esteem, were next introduced. Random assignment of students to one of the three personality scales yielded a total of seven individuals for Self-Esteem, seven for Social Participation, and eight for Tolerance. The following definitions, representing an amalgamation of definitions reported for corresponding CPI scales (Gough, 1957) and the author's as yet unpublished Jackson Personality Inventory, were presented to student item writers. (Since there is no Self-Esteem scale on the CPI, the Social Presence scale, with a substantially equivalent scale definition, was employed).

*Social Participation.* Sociable, friendly, gregarious; will eagerly join a variety of social groups, seeks both formal and informal association with others. Values positive interpersonal relationships; outgoing, sociable, participative temperament.

*Tolerance.* Broad-minded, undogmatic, open-minded; accepts people even though their beliefs and customs may differ from his own; open to new ideas; free from prejudice; permissive, accepting, and nonjudgmental in social beliefs and attitudes.

*Self-Esteem.* Self-assured, confident, self-sufficient; poised in dealing with others; not easily embarrassed or influenced by others; imperturbable in interpersonal situations; poised, spontaneous, and self-confident in personal and social interaction.

In addition to scale definitions, students were provided with a 45-minute lecture on the basics of item writing. Basic principles of simple, direct good grammatical usage; conciseness in statements; freedom from extreme levels of evaluation; the concept of content saturation

tion; and use of medium levels of item popularity were emphasized in this presentation. Students were then instructed to prepare a set of 10 true-keyed and 10 false-keyed items bearing on the personality dimension to which they were assigned. It was suggested that they not spend more than two hours at this task. As an added inducement, students were advised that the item writer in each scale category whose scale obtained the highest validity with respect to peer ratings would receive a prize of \$10.00. Students were further instructed to identify after they had written all 20 items the two weakest true-keyed and two weakest false-keyed items in their set. These were excluded, but no further editorial discretion over item selection was exercised.

The set of 22 16-item scales together with separate 16-item scales for Acquiescence and for Desirability was incorporated in a booklet entitled "Personality Inventory," and was designated Student Personality Inventory for reporting purposes.

The Desirability scale was taken from Form A of the Personality Research Form (Jackson, 1967); the Acquiescence scale was comprised of CPI items drawn randomly from the neutral range of desirability. Also included in the booklet were the three corresponding scales from the CPI. In order to conform to time limitations, and to make the CPI scales comparable in length to the students' scales, each of three sets of 16 items was chosen randomly from the CPI items comprising each of the three scales.

After the residents in the women's residence had completed the Jackson Personality Inventory and the Student Personality Inventory, they were instructed to complete a schedule containing peer ratings and self-ratings which served as the two criterion measures. The roommate ratings used for the target traits were based on a 9-point rating scale, ranging from extremely characteristic to extremely uncharacteristic of the degree to which the roommate possessed the named trait, Social Participation, Tolerance, or Self-Esteem, each with the identical definitions given item writers.

## *Results*

### *Analyses of Predictor Relationships*

Table 1 presents the intercorrelations of the Self-Esteem, Social Participation, and Tolerance scales for the scales prepared by each of the psychology students, as well as the corresponding scales taken from the JPI and the CPI. Also included in Table 1 are the KR-20 reliability coefficients for the student scales. Looking first at the Self-Esteem scales, one notes that student scales were substantially correlated with

TABLE 1

*Intercorrelations among the New Student Personality Inventory Scales Especially Constructed for this Study and JPI and CPI Scales*  
(*N* = 116)

Scale		Psychology Students							
Self-Esteem	1	1	2	3	4	5	6	7	
	2	.76	-						
	3	.55	.48	-					
	4	.73	.66	.60	-				
	5	.42	.41	.35	.47	-			
	6	.53	.44	.42	.62	.46	-		
	7	.63	.46	.59	.64	.48	.53	-	
JPI		.65	.49	.67	.67	.38	.57	.74	
CPI		.45	.46	.29	.48	.34	.23	.32	
KR-20 reliability estimate		.62	.72	.48	.74	.40	.50	.63	
		Psychology Students							
Social Participation	8	8	9	10	11	12	13	14	
	9	.42	-						
	10	.40	.75	-					
	11	.19	.23	.26	-				
	12	.30	.49	.40	.16	-			
	13	.59	.51	.42	.17	.28	-		
	14	.30	.68	.65	.38	.53	.43	-	
JPI		.65	.51	.48	.25	.37	.62	.44	
CPI		.24	.55	.62	.12	.45	.31	.57	
KR-20 reliability estimate		.48	.60	.68	-.05	.40	.50	.62	
		Psychology Students							
Tolerance	15	15	16	17	18	19	20	21	22
	16	.44	-						
	17	.24	.32	-					
	18	.51	.45	.37	-				
	19	.34	.36	.30	.47	-			
	20	.47	.34	.11	.54	.43	-		
	21	.50	.24	.29	.38	.26	.42	-	
JPI	22	.34	.39	.23	.50	.32	.24	.25	-
CPI		.58	.42	.27	.37	.15	.40	.36	.36
KR-20 reliability estimate		.13	.29	.18	.14	.20	.19	.28	.34
		.65	.45	.16	.51	.45	.28	.53	.34

each other and with the two corresponding formal personality scales, particularly the JPI. Indeed, the correlations were high enough to suggest a substantial general factor, since in a number of cases the scale intercorrelations actually exceeded the lower-bound KR-20 estimate of reliability. The same observation holds for the Social Participation and Tolerance scales except that the scale reliabilities and intercorrelations varied over a somewhat greater range of values for these scales. In no case were any of the correlations between any of the scales pro-

duced by student item writers negative. The general tenor of these findings indicates that scale definitions were communicated to student item writers, and that these definitions and item writing instructions were apparently sufficient to cause them to agree substantially on the characteristics to be measured. Even though item analytic procedures were not used, reliabilities were promising for 16-item scales. Furthermore, a review of the entire scale intercorrelation matrix for the Student Personality Inventory indicated that the different scales were mutually independent and that they formed three distinct clusters of Self-Esteem, Social Participation, and Tolerance.

### *Analyses of Criterion Relationships*

The two primary criteria used in the study were self-ratings and roommate ratings. Respective data for these two sets of criterion relationships are contained in Tables 2 and 3.

It may be noted that self-ratings were differentially sensitive to the average validities for the three scales, as they were highest for the Self-Esteem scale and lowest for the Tolerance scale. However, considering the reliabilities of the Student Personality Inventory scales and of the self-ratings, which were based on a single judge, these are indeed high correlations. For all three traits every one of the 22 Student Personality scales correlated significantly with self-ratings. It is noteworthy that the more reliable scales developed by the psychology students were superior to the less reliable ones in terms of their correlation both with self-ratings and with roommate ratings. Similarly, females showed a tendency to be superior to males in the validity of the scales that they produced, an outcome paralleling findings in the

TABLE 2  
*Validity Coefficients as a Function of Scale-Construction Strategy Relative to the Self-Rating  
Criterion Measure  
(N = 116)*

		Targeted Traits			
		SEs	Soc. P	Tol	Average
New SPI Scales	Average Psychology Student	.61	.38	.33	.44
	Most Reliable Psychology Student	.63	.44	.31	.46
	Least Reliable Psychology Student	.56	.31	.28	.38
	Average Male Psychology Student	.56	.37	.32	.41
	Average Female Psychology Student	.67	.50	.27	.48
Comparison Scales	Jackson Personality Inventory	.77	.47	.28	.51
	California Psychology Inventory	.43	.31	.19	.31

TABLE 3

*Validity Coefficients as a Function of Scale-Construction Strategy Relative to the Roommate Rating Criterion Measure*  
(*N* = 116)

		Targeted Traits			
		SEs	Soc. P	Tol	Average
New JPI Scales	Average Psychology Student	.27	.29	.18	.25
	Most Reliable Psychology Student	.30	.34	.20	.28
	Least Reliable Psychology Student	.22	.26	.16	.21
	Average Male Psychology Student	.22	.28	.16	.22
	Average Female Psychology Student	.29	.33	.20	.27
Comparison Scales	Jackson Personality Inventory	.30	.34	.23	.29
	California Psychological Inventory	.12	.07	.08	.09

social perception area indicating greater sensitivity of females to the implicit network of trait relationships (Lay, 1970).

Considering the relationships with standardized personality tests, it is noteworthy that for Self-Esteem and Social Participation the JPI did show higher relationships on the average both for self-ratings and for roommate ratings criteria than did the Student Personality Inventory. For Tolerance, the JPI was superior in validity relative to roommate ratings, but not self-ratings. The CPI did yield substantially lower validities than did the student questionnaire for both self-ratings and roommate ratings. Particularly disappointing were the relationships between the CPI items and the targeted roommate ratings, which averaged only .09. It should be recognized, of course, that the items in this comparison did not comprise the entire set of items for each CPI scale, but 16-item subscales chosen at random from the longer item set. Similarly, it should be borne in mind that each of the JPI scales was comprised of 20 items. Nevertheless, even if the longer CPI scales as compared with the shorter ones produced somewhat higher validities, this outcome indicates at least that the CPI scales are less efficient than are the student scales in predicting a criterion. If the JPI and CPI scales are taken as representative of standardized personality inventories, it is quite clear that the students in the study can generate convergent validities in a comparable range to those of standardized inventories. Considering the short length of the Student Personality Inventory scales, as well as the fact that no internal consistency or other item analytic procedures were applied, these validity coefficients compare very favorably with those appearing in the literature for published personality tests which have used these and other empirical procedures.



*Analyses of Response Styles*

Many years ago (Cronbach, 1950; Jackson and Messick, 1958), it was recognized that response styles as distinguished from the logical validity of scales might contribute to their empirical validity while detracting from the measurement of the construct which the scale was designed to assess. In an effort to evaluate the possible role of response styles in the Student Personality Inventory, scales for acquiescence and desirability, both derived from a successive intervals scaling of the California Psychological Inventory, were included in the analysis. Since Student Personality Inventory scales proved to be largely independent of scores from the Acquiescence scale, these correlations are not reported. Table 4 reports the correlations between the average scales developed by the psychology students and the Desirability scale. Also presented are the corresponding correlations for the JPI and the CPI. As might reasonably be expected, correlations between Desirability and Self-Esteem are moderately high, less so between Desirability and Social Participation, and least between Desirability and Tolerance. Of the scales reported, the only ones in which a systematic effort was made to suppress desirability variance in the total score were the scales derived from the JPI. In the construction of the JPI, the component of the total score correlated with Desirability was removed by partial regression procedures, and the residual component, uncorrelated with Desirability, used as the basis for item selection. A variant of the Differential Reliability Index (Jackson, 1967; Neill and Jackson, in press) was employed. This had the effect of subtracting the squared biserial correlation of an item with the Desirability scale from the squared item-total scale biserial, where the total scale comprised the component uncorrelated with Desirability.

TABLE 4  
*Correlations with Desirability as a Function of Scale-Construction Strategy*  
(*N* = 116)

		Targeted Traits			
		SEs	Soc. P	Tol	Average
New SPI Scales	Average Psychology Student	.44	.33	.11	.29
	Most Reliable Psychology Student	.45	.42	.11	.32
	Least Reliable Psychology Student	.43	.22	.13	.26
	Average Male Psychology Student	.48	.31	.11	.30
	Average Female Psychology Student	.43	.47	.14	.34
Comparison Scales	California Psychological Inventory	.31	.59	.62	.51
	Jackson Personality Inventory	.36	.25	.24	.28

The results from the present analysis would appear to provide support for this procedure, at least for the Self-Esteem and Social Participation scales, for which the JPI scales have shown the least high correlation with the external desirability scale. But, in the case of the SPI scales, it should be recognized that correlations with Desirability were probably not excessively high, especially when compared with those for the corresponding CPI scales. It would, of course, be appropriate to sample scales more widely, especially from the set of those with extreme levels of desirability. But it nevertheless is of some considerable importance to learn from the present study that novice item writers can develop scales relatively free of response style effects.

### *Discussion*

When Ashton and Goldberg published their findings—startling to many—that the average graduate student in psychology in two hours or less was capable of producing personality scales of equal reliability and validity to those developed by far more expensive and time-consuming External strategies, they called for additional studies to ascertain the generality of their findings to other samples of item writers, targeted traits, and subject populations. In the present study the investigator has sought to fill this requirement and, indeed, has demonstrated that undergraduate majors in psychology can produce scales about as high in validity as those prepared by the Ashton-Goldberg graduate students. The validity coefficients in the present study are even more striking when one considers that the reliability of the roommate ratings was attenuated by virtue of the use of only a single judge, albeit one well acquainted with the subject, rather than the larger number employed by Ashton and Goldberg. Findings such as those reported in the present study and by Ashton and Goldberg can not but hasten the demise of the unquestioned ascendancy of the External strategy of personality scale construction. In the past, rational or intuitive scales have been judged by some, including the writer, to be superior to External scales on the logical grounds that they yielded less ambiguous data, provided more systematic sample of relevant behaviors, and offered less susceptibility to nuisance variables. Now, when added to previous evidence (Hase and Goldberg, 1967) to the effect that predictions based on linear combinations of externally-derived scale scores were no more valid than were those based on linear combinations of intuitively-derived scale scores, new evidence exists that relatively unsophisticated psychology students can generate personality items possessing higher validities than those derived from carefully selected items based on an external criterion. With this kind

of evidence, there can be little or no justification for using the External strategy as the sole method of personality scale construction. Perhaps the only defense for using such a strategy would be in situations where one is in ignorance about the nature of the criterion. Even here, however, the present author would argue for a conceptual analysis of the criterion as an alternative to relatively blind empirical methods for discovering its components. Perhaps a further direction of research might be to examine the alternative benefits of a conceptual analysis of incompletely understood criteria as against the use of the External strategy in an exploratory context.

## REFERENCES

- Ashton, S. G. and Goldberg, L. R. In response to Jackson's challenge: The comparative validity of personality scales constructed by the external (empirical) strategy and scales developed intuitively by The experts, novices, and laymen. *Journal of Research in Personality*, 1973, 7, 1-20.
- Cronbach, L. J. Further evidence on response sets and test design. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1950, 10, 3-31.
- Goldberg, L. R. Parameters of personality inventory construction and utilization: A comparison of prediction strategies and tactics. *Multivariate Behavioral Research Monograph*, 1972, No. 72-2.
- Gough, H. G. *Manual for the California Psychological Inventory*. Palo Alto, Calif.: Consulting Psychologists Press, 1957.
- Hase, H. D. and Goldberg, L. R. Comparative validity of different strategies of constructing personality inventory scales. *Psychological Bulletin*, 1967, 67, 231-248.
- Jackson, D. N. *Manual for the Personality Research Form*. Goshen, New York: Research Psychologists Press, 1967.
- Jackson, D. N. The dynamics of structured personality tests: 1971. *Psychological Review*, 1971, 78, 229-248.
- Jackson, D. N. and Messick, S. Content and style in personality assessment. *Psychological Bulletin*, 1958, 55, 243-252.
- Lay, C. H. Trait inferential relationships and judgments about the personality of others. *Canadian Journal of Behavioural Science*, 1970, 2, 1-17.
- Meehl, P. E. The dynamics of "structured" personality tests. *Journal of Clinical Psychology*, 1945, 1, 296-303.
- Neill, J. A. and Jackson, D. N. Minimum redundancy item analysis. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1976, in press.

## IMPROVING THE VALIDITY OF AFFECTIVE SELF-REPORT MEASURES THROUGH CONSTRUCTING PERSONALITY SCALES UNCONFOUNDED WITH SOCIAL DESIRABILITY: A STUDY OF THE PERSONALITY RESEARCH FORM<sup>1</sup>

ROBERT D. ABBOTT

California State University, Fullerton

Jackson's Personality Research Form (PRF) was investigated at the item and scale level with respect to the degree to which responses are confounded by social desirability. Jackson's usage of the differential validity index resulted in large proportions of items neutral in social desirability and scales relatively balanced in the number of items keyed for socially desirable and socially undesirable responses. These item and scale characteristics, the low correlations of PRF trait scales with social desirability scale scores, and the results of a component analysis of PRF trait scales and social desirability scales supported the discriminant construct validity of the PRF trait scales with respect to social desirability.

JACKSON (1967) has recently introduced the Personality Research Form (PRF). Forms AA and BB of the PRF are parallel forms providing scores on 20 trait scales and on two stylistic scales. The 20 trait scales were based upon Murray's needs and the two stylistic scales were designed to provide measures of nonpurposive responding and social desirability. Jackson's inclusion of a desirability scale in the PRF is but one indication of his efforts to reduce the confounding of scores on PRF scales with the tendency to respond in a socially desirable manner (Edwards, 1957, 1970). Column 1 of Table 1 shows the distribution of the absolute values of the correlations of the PRF trait scales with the PRF Desirability (*Dy*) scale. A comparison of

<sup>1</sup> This is a version of a paper presented to the Western Psychological Association, Los Angeles, California, April, 1970.



these values with the correlations between Social Desirability (*SD*) scales and scales in other inventories such as the California Psychological Inventory and the Minnesota Multiphasic Personality Inventory (MMPI) (Abbott, 1971; Edwards, 1970) has shown that the PRF scales are much less confounded with social desirability than are scales in other inventories. Jackson (1967, 1970) reported achieving this goal by using items with high "content" saturation with Jackson and Messick's (Jackson, 1970) differential validity index furnishing a quantitative measure of content saturation.

The purpose of the present paper was to present a series of analyses of PRF items and scales to determine the apparent effects of the use of the differential validity index on more traditional item psychometric indices which have been used to reduce the confounding of trait scales and social desirability. Such information could be of potential value to research workers in personality test development who wish to employ a methodology based upon the differential validity index for enhancing the construct validity of a scale by minimizing response set confounding.

### *Method*

Following directions reproduced in Edwards (1970), Group 1, consisting of 100 students, rated the Social Desirability Scale Values (SDSV) of the 440 items in Form AA of the PRF. The mean SDSV was obtained for each item.

As part of an independent test research project, Group 2 (109 males and 109 females) followed self-description instructions and responded to the items in Form AA of the PRF, to items from the Edwards (1957) MMPI Social Desirability (*SD*) scale, and to items in Welsh's Repression (*R*) scale.

### *Results and Discussion*

#### *Item Level*

Edwards (1957) proposed that one way to reduce the effects of social desirability on responses to personality scales would be to use items which have SDSVs in the middle or neutral range of the SDSV continuum. For large, relatively unselected, groups of personality items and constructs, the distribution of SDSVs has been shown (Cruse, 1965; Edwards, 1966) to be bimodal with the modes falling somewhere around 3 and 7 on the 9 point SDSV continuum. Figure 1 shows the distribution of the SDSVs of the 400 PRF trait items, and the distributions of SDSVs reported by Edwards (1966) and Cruse (1965).



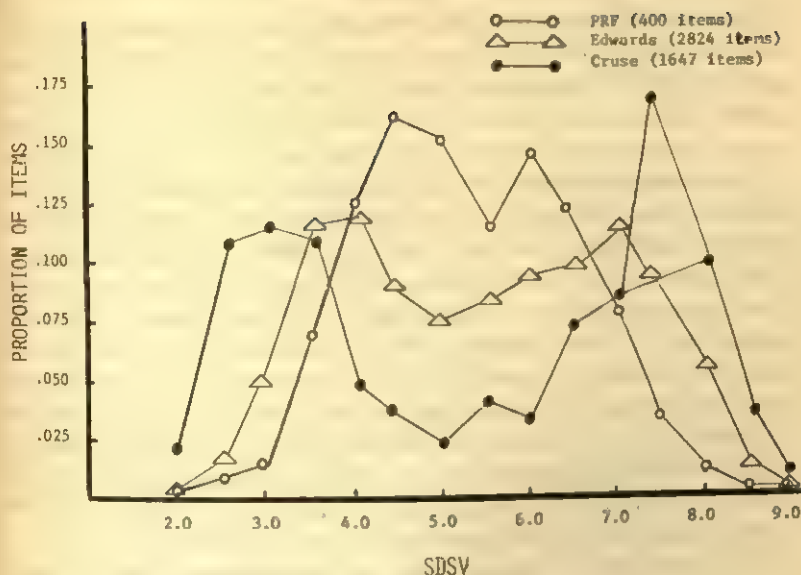


Figure 1. Distributions of SDSVs from three item pools.

Figure 1 clearly shows that the proportion of items in the PRF with neutral SDSVs is larger than the proportion of items with neutral SDSVs in the Cruse (1965) list of personal constructs or in the Edwards (1966) list of items from which he constructed his Edwards Personality Inventory (Edwards, 1967). The usage of the differential validity index by Jackson has resulted in items which are relatively neutral in social desirability. Given the generalizability (Edwards, 1970) of SDSV mean ratings, the PRF items would probably be rated in much the same way by other groups.

### Scale Level

Columns 2 and 3 of Table 1 show the frequency distribution of the absolute correlations of the 20 PRF trait scales with the Edwards *SD*

TABLE 1

Frequency Distribution of Absolute Correlations of the 20 PRF Trait Scales with Three Desirability Scales

	$r_{py}$	$r_{sd}$	$r_{bsd}$
.50-.59	0	0	0
.40-.49	0	1	0
.30-.39	4	4	0
.20-.29	5	5	7
.10-.19	7	3	6
.00-.09	4	7	7

(Edwards, 1957) scale and the Balanced True-False (BSD) Social Desirability scale that Edwards and Abbott (Edwards, 1970) constructed from the MMPI. In no case did any of the social desirability scales account for more than 20% of the variance in a PRF trait scale. These results strongly replicate those of Jackson (1967) with his *Dy* scale.

To investigate further the relationship between *SD* scale score and scores on the PRF scales, scores on the 22 PRF scales and Edwards *SD* scale were intercorrelated, factor analyzed by the principal component method, and the six factors with eigenvalues greater than one were rotated by Kaiser's varimax method. Table 2 presents all rotated normalized loadings greater than |.40| and the communality ( $h^2$ ) of each of the scales. It is seen that *Dy*, Jackson's measure of desirability, and *SD*, Edwards measure, both marked factor V, which accounted for less than 25% of any of the 20 PRF trait scale's common variance. This evidence not only shows that Jackson has succeeded in his attempt to minimize the importance of social desirability in the PRF but also strongly replicates research reported in his manual.

Earlier studies by Edwards and his co-workers (Edwards, Diers, and Walker, 1962; Edwards and Walsh, 1963) have shown that several

TABLE 2  
*Rotated Normalized Factor Loading Matrix of the PRF and Marker Scales*

Scales	I	II	III	IV	V	VI	$h^2$
<i>Ab</i>			43	-76			67
<i>Ac</i>		91					75
<i>Af</i>			90		-41		68
<i>Ag</i>				86	50		76
<i>Au</i>	47		-80				56
<i>Ch</i>	86	50					75
<i>Cs</i>	-99						65
<i>De</i>				81	48		71
<i>Do</i>				81	-43		66
<i>En</i>		87					67
<i>Ex</i>				86			54
<i>Ha</i>	-80						75
<i>Im</i>	95						69
<i>Nu</i>			83				63
<i>Or</i>	-95						68
<i>Pl</i>	53	-51	49	43			51
<i>Se</i>	50	76					63
<i>Sr</i>	-41		60	59			70
<i>Su</i>			86				60
<i>Un</i>		91					67
<i>In</i>						96	72
<i>Dy</i>					-89		72
<i>SD</i>					-98		72

Note—Only loadings greater in magnitude than .40 are shown. Decimals have been omitted.

TABLE 3  
*Distribution of SDSV's, Correlation with the Edwards SD Scale ( $r_{SD}$ ), Proportion of Neutral Items  $P(N)$ , and Imbalance in SD-SUD Keying for Each of the 22 PRF Scales*

SDSV	Ab	Ac	Af	Ag	Au	Ch	Cs	De	Do	En	Ex	Ha	Im	Nu	Or	Pl	Se	Sr	Su	Un	In	Dy
8.50-8.99																						1
8.00-8.49																						
7.50-7.99			3	2	2	2		1	1	4	1		1			2	2	1		2	3	3
7.00-7.49		1	4	3	2	4		2	1	4		1	3	5	1		7	2	1	4	3	2
6.50-6.99	1	1	1	1	4	3	4	2	1	4	1	1	6		3	3	1	4	3	3	3	4
6.00-6.49	4	2	1	1	4	2	3	5	4	1	1	5	2		2	6		4	4			
5.50-5.99	1	4	1	2	4	2	3	2	3	3	5	1	2		2	3	3	2	5	3		
5.00-5.49	3	1	1	1	1	4	3	3	7	4	5	7	2		3	3	5	4	4	2	1	1
4.50-4.99	2	2			4	2	3	3	3	2	4	3	4	1	8	3	3	3	2			
4.00-4.49	4	2	2	2	2	2	4	3	3	2	4	3	2	5	2	2	2	3	1	1	3	3
3.50-3.99	3	3	1	1	2	5	3	3	1	5	2	3	2	4	2	1	3			4	2	4
3.00-3.49		3	4	5	1			1			2									1	1	2
2.50-2.99	2	1		2																		
2.00-2.49			2																			
1.50-1.99																						
1.00-1.49																						
$r_{SD}$	-.08	.19	.28	-.26	-.03	.00	.00	-.28	.27	.25	.08	-.12	-.09	.22	.14	.04	.11	-.16	-.21	.11	-.11	.57
$P(N)$	.50	.40	.20	.25	.55	.50	.65	.65	.85	.35	.75	.80	.70	.05	.80	.75	.30	.65	.75	.40	.20	.05
Imbalance	-.15	.35	.45	-.40	-.05	.20	.10	-.30	.15	.35	.25	-.35	-.10	.45	.35	.15	.50	.00	-.05	.40	-.40	.50

scale psychometric indices derived from the item SDSVs are predictive of the correlation of a scale with the *SD* scale and are thus helpful in predicting and interpreting the degree to which trait scale scores are confounded with the tendency to respond in a socially desirable direction. These indices have included the imbalance (IMB) in the Social Desirability-Social Undesirability (SD-SUD) keying, to be defined shortly, and the proportion of items in a scale with neutral SDSVs,  $P(N)$ . For MMPI scales, as IMB increases, the magnitude of the correlation with the *SD* scale increases, and as  $P(N)$  increases, the correlation with the *SD* scale decreases.

For each PRF scale, Table 3 shows the distribution of SDSVs of the items in each scale, the correlation of the PRF scale with the *SD* scale, the imbalance in the PRF scale's SD-SUD keying computed by subtracting .5 from the proportion of items in a scale keyed for a socially desirable response, as well as, the proportion of items in a scale with SDSVs between 4 and 6 on the 9-point SDSV continuum,  $P(N)$ . These item characteristics provided much information about the correlation of trait scale scores with scores on the *SD* scale.  $P(N)$  was correlated  $-.460$  with the correlation of a scale with the *SD* scale. Thus as the proportion of neutral items in a PRF scale increased, the correlation with the *SD* scale decreased. The degree of imbalance in SD-SUD keying correlated  $.48$  with the correlation of a scale with the *SD* scale. Thus as the imbalance increased, the correlation of the scale with the *SD* scale increased. These findings, which replicate those with the MMPI, extend the usefulness of these item characteristics to items of nonpathological content taken from trait scales that have been designed to measure individual differences in "normal" personality traits.

Jackson's use of the differential validity index and of item selection techniques has resulted in scales consisting of a greater proportion of neutral items, as well as in scales which, in general, have a balance in their SD-SUD keying. This study has indicated that PRF scales that do have correlations with the *SD* scale are those with smaller proportions of neutral items or with imbalances in their SD-SUD keying, i.e., the *Dy* scale. However, the relatively small confounding of variance in the PRF trait scales with the *SD* scale has lent support to the discriminant validity of the PRF trait scales.

#### REFERENCES

- Abbott, R. D. A factor analysis of the California Psychological Inventory and Edwards Personality Inventory. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1971, 31, 549-553.
- Cruse, D. B. Social desirability scale values of personal concepts. *Journal of Applied Psychology*, 1965, 49, 342-344.

- Edwards, A. L. *The social desirability variable in personality assessment and research*. New York: Holt, 1957.
- Edwards, A. L. Relationship between probability of endorsement and social desirability scale value for a set of 2,824 personality statements. *Journal of Applied Psychology*, 1966, 50, 238-239.
- Edwards, A. L. *Manual for the Edwards Personality Inventory*. Chicago: Science Research Associates, 1967.
- Edwards, A. L. *The measurement of personality traits by scales and inventories*. New York: Holt, 1970.
- Edwards, A. L., Diers, C. J., and Walker, J. N. Response sets and factor loadings on sixty-one personality scales. *Journal of Applied Psychology*, 1962, 46, 220-225.
- Edwards, A. L. and Walsh, J. A. The relationship between the intensity of the social desirability keying of a scale and the correlation of the scale with the Edwards' *SD* scale and the first factor loading of the scale. *Journal of Clinical Psychology*, 1963, 19, 200-203.
- Jackson, D. N. *Manual for the Personality Research Form*. Goshen, N. Y.: Research Psychologists Press, 1967.
- Jackson, D. N. A sequential system for personality scale development. In C. Spielberger (Ed.) *Current topics in clinical and community psychology* (Vol. 2). New York: Academic Press, 1970.





## THE RELIABILITY AND VALIDITY OF TWO OBJECTIVE MEASURES OF ACHIEVEMENT MOTIVATION FOR ADOLESCENT FEMALES

MICHAEL POMERANTZ

University of Connecticut

CHARLES B. SCHULTZ

Trinity College

Two measures of motive to succeed were revised for administration to female ninth grade students ( $N = 71$ ). Scores on Hermans' (1970) Prestatie Motivatie Test (PMT) yielded a high degree of internal consistency, comparable to that obtained with males, which was greater than that found for scores on the present version of Mehrabian's (1969) Resultant Achievement Motivation (RAM) Test. In separate validation analyses, scores on the PMT were observed to correlate positively and substantially with each of two measures of school achievement and with questionnaire data on school-related attitudes and behavior. Although the correlations of the RAM scores with achievement measures were in the same direction, they were weaker than those for the PMT scores. The relationships between the two measures of motive to succeed and various internal causal ascriptions were different and low for the two instruments.

In order to apply achievement motivation theory to educational problems, an objective measure of motive to succeed ( $M_s$ ) is needed which can be conveniently administered to groups of adolescents. Even more important, the instrument must possess reasonable and comparable reliability and validity for both males and females. Two sources of weakness of current projective tests and, in particular, of many objective measures that are intended to represent the same characteristics are their lack of reliability and their absence of high correlations with each other (Weinstein, 1969). One implication of

The authors gratefully acknowledge the assistance of Charles Clock, Edward Bjerman, and Francis Whittle of the West Hartford School District for their cooperation in the conduct of this research.

Copyright © 1975 by Frederic Kuder

these findings is that objective tests may measure different aspects of the same construct.

Some success has been reported with newer instruments not evaluated by Weinstein (1969). One is Mehrabian's (1968, 1969) Resultant Achievement Motivation (RAM) Test which has male (RAMm) and female (RAMf) subscales. Several studies present evidence for the validity of the subscales (Cohen, Reid, and Boothroyd, 1973; Farley and Mealiea, 1973; Mehrabian, 1968, 1969; Raffini and Rosemier, 1972; Reid and Cohen, 1973, 1974; Weiner and Potepan, 1970). Although Mehrabian has considered the RAM to be a resultant measure of the *Ms* minus the motive to avoid failure (*Maf*), there has been some speculation that it may measure the *Ms* alone (Weiner and Potepan, 1970). In the present study both propositions are considered. A second promising *Ms* instrument is Hermans' (1970) Prestatie Motivatie Test (PMT) which though not designed for use with females has been shown to compare favorably with the RAMm subscale when male subjects were employed (Schultz and Pomerantz, 1974).

The purpose of this study was to alter existing instruments so that a combined form of the RAMm and RAMf could be administered simultaneously to a mixed male and female population and so that both instruments could be given to a younger and more heterogeneous group of subjects than that represented by college students. The instruments were analyzed for reliability; validity was assessed by correlating scores on each of them with academic performance, with a measure of locus of control, and with scores derived from items on a school attitude and behavior questionnaire reflecting variables such as educational aspirations and frequency of doing homework.

### *Method*

One hundred and sixty-four subjects from two suburban junior high schools were randomly drawn from a large pool of ninth grade students and tested. Of these, 71 females were included in the following analyses. Two subjects who had not completed both test batteries were dropped from the analyses.

Two batteries of tests were administered as a part of another project reported in more detail elsewhere (Schultz and Pomerantz, 1974). Several instruments were modified to improve their readability and to make them more relevant to a younger age group and to a school environment.<sup>1</sup> These included the following: Mehrabian's (1969) female

<sup>1</sup> It should be noted that 12 items in the RAMf were altered to change either the age or sex orientation so that they could be administered to both sexes simultaneously. Revised versions of the RAM, PMT, and DAS are available from Charles B. Schultz, Department of Education, Trinity College, Hartford, Connecticut 06106.

scale of the Resultant Achievement Motivation Test (RAMf), which was administered with the male scale but analyzed separately, Hermans' (1970) Prestatie Motivatie Test (PMT), and the Debilitating Anxiety Subscale (DAS) of the Achievement Anxiety Scale (Alpert and Haber, 1960). The Intellectual Achievement Responsibility (IAR) questionnaire (Crandall, Katkovsky, and Crandall, 1965) was employed in its original form as a measure of locus of control. This instrument was divided into the following four subscales of internal causal ascriptions: Success to ability, success to effort, failure to lack of ability, and failure to lack of effort.<sup>2</sup> The subjects also responded to a questionnaire of Likert scale items related to school attitudes and behavior which reflect success striving, persistence, and achievement behavior. (See Table 3.) The Comprehensive Test of Basic Skills (CTBS) served as one index of school achievement.

Test batteries were administered in two sessions to groups of approximately 25 subjects. The RAMf, DAS, and the school attitude and behavior questionnaire were completed during the first testing period and the PMT and IAR were completed approximately two weeks later during a second session. All the items except the school attitude and behavior questionnaire were presented via 35 mm. slides projected onto a screen. The subjects read each item while simultaneously hearing a tape-recorded reading of the item which was synchronized to the slide projector. The CTBS was administered approximately two months earlier by school personnel independently of this investigation.

For each subject, three resultant achievement motivation scores were obtained. The first was the RAMf score which Mehrabian (1969) described as a resultant index of *Ms-Maf*. Since Weiner and Potepan (1970) suggested that the RAM may measure *Ms* alone, a second resultant was computed by subtracting the DAS z-score for each subject from her RAMf z-score; this difference score was labelled Mehrabian's resultant (RAMf-DAS). The third resultant score was similarly computed by subtracting the DAS z-score for each subject from her PMT z-score. This difference score was called Hermans' resultant (PMT-DAS).

### *Results and Their Interpretation*

Both *Ms* measures were analyzed for their internal consistency (Cronbach, 1951). This analysis yielded an alpha coefficient of .59 for

<sup>2</sup> Weiner (e.g., Weiner and Potepan, 1970) divided his adult version of the IAR into subscales of ability and effort or motivation. Crandall categorized all items other than those which reflect effort as simply undifferentiated, since these items may refer to more than ability (personal communication). Crandall's subscales were employed in the present analyses with the labels used by Weiner and Potepan for the sake of consistency.

the RAMf and .84 for the PMT. The greater consistency index for the PMT as compared with that for the RAMf approximates the coefficient of .82 obtained by Hermans (1970). These results are comparable to other findings with males in which the alpha coefficient for the RAMm was .55 and the alpha coefficient for the PMT was .91 (Schultz and Pomerantz, 1974).

Resultant achievement motivation is presumed to be positively related to measures of academic achievement (Hermans, 1970). Accordingly, the RAMf, RAMf-DAS, and PMT-DAS scores were correlated with each of two measures of school achievement, teachers' grades (GPA), and scores on a standardized test (CTBS). (See Table 1.)

Scores on the PMT exhibited a substantial relationship to academic achievement for females and compared favorably with similar correlations for males (Schultz and Pomerantz, 1974). When the resultant PMT-DAS score was correlated with achievement, the relationship was somewhat stronger than that for the PMT alone. Both relationships would appear to support the use of the PMT as an index of *Ms*. In contrast, the relatively low correlations between scores on the RAMf and school achievement would suggest that the RAMf has mild predictive validity whether conceived of as a measure of *Ms* or of resultant achievement motivation. Standing on the DAS was related in a predictably negative manner to level of achievement. However, a correlation no higher than that for the DAS was obtained from a combination of the two measures in the RAMf-DAS resultant. Since the RAMf alone was not related to academic achievement and in combination with the DAS did not add to what was obtained from use of the DAS alone, the RAMf would appear to lack validity as a measure of *Ms* or of *Ms-Maf*. Furthermore, since the PMT-DAS and the

TABLE 1  
*Correlations of Achievement Motivation Measures with Indices of School Achievement*  
(*N* = 69)

Measures of Achievement Motivation	Indices of School Achievement	
	GPA	CTBS
PMT	.56***	.50***
PMT-DAS	.60***	.61***
RAMf	.11	.24*
RAMf-DAS	.40**	.55***
DAS	-.44***	-.52***

\*  $p < .05$ .

\*\*  $p < .01$ .

\*\*\*  $p < .001$ .



RAMf variables were nonsignificantly related ( $r = .08$ ), they, therefore, could not be considered equivalent resultant indices.

Measures of achievement motivation have been validated by relating them to causal ascriptions for success and failure (Cohen, et al., 1973; Weiner and Potepan, 1970). Achievement needs are presumed to be positively related to the attribution of success to ability and to effort as well as to the attribution of failure to lack of effort. Achievement needs are also presumed to be inversely related to the attribution of failure to lack of ability (Weiner and Potepan, 1970). Table 2 summarizes these relationships in the present study. Few of the predicted correlations were obtained. The PMT and the PMT-DAS variables were positively related to a significant degree ( $p < .05$ ) to the attribution of success to ability, and the inverse relationship between the RAMf-DAS variable and the attribution of failure to lack of ability was statistically significant ( $p < .05$ ).

The case for the validity of the PMT with the female adolescents of the present study rested largely on its correlation with attribution of success to ability. In this respect, the findings were consistent with both theoretical expectations and the most frequent previous findings. The case for the validity of the RAMf rests largely on its negative correlation with the attribution of failure to lack of ability. This relationship replicated earlier findings with undergraduate males and females (Weiner and Potepan, 1970) and with undergraduate males (Cohen et al., 1973). However, Cohen et al. (1973) failed to obtain a similar relationship with undergraduate females, and Schultz and Pomerantz (1974) failed to find it with adolescent males. Neither measure of *Ms* was correlated positively with the attribution of failure to lack of effort. To the knowledge of the writers, this relationship has not been reported in any of the other studies in which measures of *Ms* were correlated with the subscales of the IAR when slightly different populations had been used. The present findings may reflect difficulties

TABLE 2  
*Correlations of Achievement Motivation Measures  
with Internal Causal Ascriptions (N = 69)*

Internal Causal Ascriptions	Measures of Achievement Motivation			
	PMT	PMT-DAS	RAMf	RAMf-DAS
Success to Ability	.29*	.29*	.03	.16
Success to Effort	.12	.07	.13	.09
Total Internal for Success	.26*	.22	.10	.16
Failure to Lack of Ability	.08	-.04	-.19	-.25*
Failure to Lack of Effort	.03	.13	-.04	.10
Total Internal for Failure	.07	.06	-.14	-.07

\* $p < .05$ .

with the achievement motivation or attribution models as much as with the invalidity of the measuring devices.

The validity of measures of achievement motivation has been assessed by relating them to questionnaires measuring achievement-related attitudes and behaviors (Mehrabian, 1969). Table 3 provides information on self-reports of academic achievement attitudes and behaviors. Most of the correlations which were significant were in the predicted, positive direction. Students scoring high on the various achievement motivation measures, as compared with those who earned low scores, tended to report higher educational aspirations, to place a greater value on grades, to indicate that their grades reflected their knowledge, and to report having higher grades and doing more homework. Among the measures of achievement motivation, the correlations were clearly highest when Hermans' PMT was used as a measure of *Ms*. Generally the RAMf-DAS correlations were stronger than were those for the RAMf alone.

Mehrabian originally designed the RAMf as a resultant measure of *Ms-Maf*. This instrument as currently modified for use with adolescent females was nonsignificantly related to the DAS ( $r = .03$ ). If the RAMf was functioning as a resultant, it should be inversely related to debilitating anxiety because that factor is by definition a component of the resultant. Furthermore, the RAMf was only weakly related to the CTBS and was not significantly related to grades or to internal causal ascriptions. Alternatively, the RAMf might be considered a measure of the *Ms* alone as suggested by the low correlation with the measure of test anxiety. Across significant and nonsignificant results, there was a weak trend for the RAMf-DAS, rather than for the RAMf alone, to

TABLE 3  
*Correlations of Achievement Motivation Measures with Selected Items  
from the School Attitude and Behavior Questionnaire (N = 66\*)*

Self-Reported School Attitudes and Behaviors	Measures of Achievement Motivation			
	PMT	PMT-DAS	RAMf	RAMf-DAS
Educational Aspirations	.33**	.22	.29*	.25*
Importance of Grades	.51***	.43***	.07	.21
Extent to Which Grades Reflect Knowledge	.47***	.39**	.11	.21
Usual Grades	.51***	.56***	.09	.38**
Amount of Homework Done	.57***	.46***	-.02	.13
Frequency of Doing Homework	.43***	.33**	.08	.14

\* Three additional students were omitted from this analysis because they did not complete all items on the questionnaire.

\*  $p < .05$ .

\*\*  $p < .01$ .

\*\*\*  $p < .001$ .

relate more strongly with questionnaire items and internal causal ascriptions. Regardless of what the RAMf is considered to be, it did possess less validity than did the PMT for use with female adolescents as revised for this experiment.

The literature on achievement motivation has made it clear that instruments which attain a measure of success in assessing achievement needs for males have not typically worked for females. This difference may have been due to the different acculturation that males and females experience. One effect of different acculturations may be to render other motivations such as fear of success more important for females than for males. Therefore, measures of achievement needs may have met with difficulty. According to this line of reasoning, such instruments are used to measure something which may exist in small quantities, if at all. However, it is also possible that, although both males and females may acquire achievement needs, these needs are expressed differently because of different socialization processes and therefore must be measured differently.

Some of the more recent studies suggest that the achievement motivation construct may be applicable to both sexes when measured by the Thematic Apperception Test (Simons and Bibb, 1974; Ollendick, 1974). Alper (1974) has speculated that because males may have shown less interest in being achievers and females may have recognized that achievement is an appropriate trait for the female role, previous findings of male and female differences in achievement motivation may have been mitigated. Although the present findings did not bear directly on these attitude changes, they were consistent with Alper's (1974) suggestions. They appeared to extend the recent trend of obtaining predicted effects with females on projective devices by yielding similar outcomes on an objective instrument.

This result in particular was the case for the PMT which has emerged from the present analyses as a relatively reliable and valid objective measure of *Ms*, at least with a school-age female population. Moreover, to the extent that the validity of the PMT is sustained, it appears that male and female achievement needs can be measured in the same way.

## REFERENCES

- Alper, T. G. Achievement motivation in college women: A now-you-see-it-now-you-don't phenomenon. *American Psychologist*, 1974, 29, 195-203.
- Alpert, R. and Haber, R. N. Anxiety in academic achievement situations. *Journal of Abnormal and Social Psychology*, 1960, 61, 207-215.
- Cohen, L., Reid, I., and Boothroyd, K. Validation of the Mehrabian

- need for achievement scale with college of education students. *British Journal of Educational Psychology*, 1973, 43, 269-278.
- Crandall, V. C., Katkovsky, W., and Crandall, V. J. Children's beliefs in their own control of reinforcements in intellectual-academic achievement situations. *Child Development*, 1965, 36, 91-109.
- Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, 16, 297-334.
- Farley, F. H. and Mealiea, W. L. Motivation and the recall of completed and incompleting achievement items. *Journal of Educational Research*, 1973, 66, 302-306.
- Hermans, H. J. M. A questionnaire measure of achievement motivation. *Journal of Applied Psychology*, 1970, 54, 353-363.
- Mehrabian, A. Male and female scales of the tendency to succeed. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1968, 28, 493-502.
- Mehrabian, A. Measures of achieving tendency. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1969, 29, 445-451.
- Ollendick, T. H. Level of *n*-achievement and persistence behavior in children. *Developmental Psychology*, 1974, 10, 457.
- Raffini, J. and Rosemier, R. Effect of resultant achievement motivation on postexam error-correcting performance. *Journal of Educational Psychology*, 1972, 63, 281-286.
- Reid, I. and Cohen, L. Achievement orientation, intellectual achievement responsibility, and the choice between degree and certificate courses in colleges of education. *British Journal of Educational Psychology*, 1973, 43, 63-66.
- Reid, I. and Cohen, L. Male and female achievement orientation and intellectual responsibility: A British validation study. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1974, 34, 379-382.
- Schultz, C. B. and Pomerantz, M. Some problems in the application of achievement motivation to education: The assessment of motive to succeed and probability of success. *Journal of Educational Psychology*, 1974, 66, 599-608.
- Simons, R. H. and Bibb, J. J. Achievement motivation, test anxiety, and underachievement in the elementary school. *Journal of Educational Research*, 1974, 67, 366-369.
- Weiner, B. and Potepan, P. A. Personality characteristics and affective reactions toward exams of superior and failing college students. *Journal of Educational Psychology*, 1970, 61, 144-151.
- Weinstein, M. S. Achievement motivation and risk preference. *Journal of Personality and Social Psychology*, 1969, 13, 153-172.



## PREDICTION OF COLLEGE ACHIEVEMENT USING THE NEED ACHIEVEMENT SCALE FROM THE EDWARDS PERSONAL PREFERENCE SCHEDULE<sup>1</sup>

RONALD R. MORGAN

Loyola University of Chicago

The purpose of this study was to investigate the utility of the Edwards need achievement scale (*n-Ach*) for predicting achievement performance, as (a) a supplement to academic aptitude test, and (b) a predictor of over- and under-achievement. Subjects were 217 college students enrolled in five sections of a general introductory psychology course. A correlational analysis was carried out among the following measures: Edwards *n-Ach* score, American College Testing Program Examination (ACT) score, overall grade point average (GPA), psychological course grade, and derived measures of over- and under-achievement.

From the results of the study, the following conclusions were drawn:

- a. Little support for the use of the Edwards *n-Ach* scale as a supplement to ability test scores in the prediction of academic performance was offered.
- b. The *n-Ach* scale was of little value in differentiating between over- and under-achievers.
- c. Further investigation is needed to evaluate a single course grade as an alternative to overall GPA as a suitable criterion of academic achievement.

THE applied psychologist is continually faced with the issue of the prediction of scholastic success in academic institutions. Scholastic aptitude and intelligence tests have proved quite useful for this prediction. However, predictions based on these measures are not perfect. In fact, they account for less than half of the variance in academic performance.

<sup>1</sup> Request for reprints should be sent to Ronald R. Morgan, Loyola University of Chicago, 820 North Michigan Avenue, Chicago, Illinois 60611.  
Copyright © 1975 by Frederic Kuder



Recent research interest in personality variables and in other non-intellectual factors has pointed to these elements as additional sources of variance in the prediction of academic achievement. The Edwards Personal Preference Schedule (EPPS) measures personality factors or "needs," such as, Achievement and Endurance, which are logically related to academic performance. Therefore, the EPPS would appear to be useful in measuring personality variables in their relationship to academic success (Edwards, 1959).

The practical value of a personality measure in the prediction of academic performance depends upon its ability to account for a portion of the criterion variance not predicted by an academic ability test. For example, Weiss, Wertheimer, and Groesbeck (1959) added *n-Ach* scores to academic aptitude test scores in a multiple regression equation. With a sample of 49 undergraduate psychology students they found the coefficient of correlation with overall grade point average (GPA) was increased from .55 to .64. In a different approach to the same problem, Goodstein and Heilbrum (1962) obtained a .24 coefficient of correlation between *n-Ach* scores and GPA after correlations due to difference in academic aptitude were partialled out ( $N = 206$ ). Gebhart and Hoyt (1958) found *n-Ach* scores to be higher for "over-achievers" than for "under-achievers." "Over-achievers" were defined as students whose grades were substantially higher than were those predicted by academic aptitude test scores; the opposite condition defined "under-achiever." In a similar study, Krug (1959) confirmed the findings.

The just described results are relevant to the construct validity of the Edwards *n-Ach* scale. Scores on the *n-Ach* scale should be positively correlated with academic achievement, when academic aptitude is held constant. A valid measure of *n-Ach* would be expected to have a significant degree of correlation with the amount of over-achievement. Achievement motivation as reflected in *n-Ach* scores, apart from intellectual ability or measured academic aptitude, would be anticipated to influence academic performance in a positive manner.

### *Purposes*

The two major purposes of this study were to provide additional evidence concerning the efficiency of the Edwards *n-Ach* measure as a supplement to standard tests of academic aptitude in predicting academic achievement and to discriminate between over- and under-achievers. This study was essentially a replication of an earlier one done by Bachman (1964). Bachman found no increment in prediction of GPA when *n-Ach* scores were added to scores on a scholastic ap-

titude test in a multiple regression equation and little success in predicting over- and under-achievement from *n-Ach* scores.

A subsidiary purpose was to examine an alternative to overall GPA as the criterion for studies of academic achievement. The development of adequate criteria is an endless problem in the study of academic achievement. The most frequently used and most often criticized criterion is the student's grade point average, taken over one or more semesters. The level of difficulty in different courses, different standards applied by each instructor in his method of evaluation, and variation in courses under different instructors all lead to the incorporation of unwanted variance in the criterion.

These difficulties might be avoided by using grades assigned by one teacher in a single course. This study employed such an approach through using introductory psychology unit examination scores as the alternative criterion of academic achievement.

It was hypothesized that

- a. The Edwards need achievement scale (*n-Ach*) would be a useful supplement to academic aptitude test scores. The multiple correlation (involving a weighted combination of the ACT and *n-Ach* variables) with GPA would be greater than the zero-order correlation between scores in the ACT and GPA earned.
- b. The Edwards *n-Ach* scale would discriminate between over- and under-achievers. A valid measure of *n-Ach* would be expected to have a significant level of correlation with the degree of over-achievement.
- c. As a criterion of achievement, a variable of obtained grades in the general psychology course would yield higher correlations with individual predictors and composites of predictors than would overall grade point average (GPA).

### Method

#### Subjects

Full-time college students enrolled in five introductory psychology sections provided data for the study. The total number of students enrolled in these sections was 290. Thirty-four of these had not taken the ACT and 28 either failed to take the EPPS or withdrew from class. The remaining 217 students (135 males, 82 females) made up the total number of subjects used in the study.

#### Measures

In addition to the previously cited measures of *n-Ach* from the EPPS, the ACT, overall GPA, and introductory psychology unit ex-

amination scores (based on mean  $T$  scores for three unit examinations in General Psychology) ( $Psy.$ ), a measure of over-achievement ( $AI_{GPA}$ ) was developed by subtracting predicted GPA from obtained GPA. The predictions employed the ACT regression equation from the ACT Research Service Report (ACT, 1965). Thus,  $AI_{GPA} = \text{obtained GPA} - \text{predicted GPA}$ . The regression equation used was as follows:

$$\text{Predicted GPA} = 0.376 + 0.046 \text{ ACT}_{ENG.}$$

$$+ 0.025 \text{ ACT}_{MATH} + 0.028 \text{ ACT}_{SOC.S} - 0.011 \text{ ACT}_{N.SCI.}$$

A similar measure of over-achievement ( $AI_{Psy.}$ ) was developed for psychology course performance. This entailed calculation of a regression equation for predicting psychology test performance from the weighted combination of ACT battery scores. The resulting equation was as follows:

$$\text{Predicted Psy.} = 31.66 + 0.0031 \text{ ACT}_{ENG.} + 0.3930 \text{ ACT}_{MATH}$$

$$+ 0.6143 \text{ ACT}_{SOC.S.} - 0.1528 \text{ ACT}_{N.SCI} \text{ Mean } T \text{ Score}$$

### *Statistical Analysis*

Product-moment coefficients of correlation were computed among the measures just listed. In addition, ACT and  $n\text{-Ach}$  scores were combined in multiple correlations with each of the two criteria of academic performance (GPA and  $Psy.$ ). Partial  $r$ 's were computed for  $r(AI_{GPA})$  ( $ACT_{Composite}$ ) with  $n\text{-Arch}$  removed and for  $r(AI_{Psy.})$  ( $ACT_{Composite}$ ) with  $n\text{-Ach}$  removed.

## *Results*

### *Prediction of Academic Performance*

In Table 1 the correlations among predictors and criteria of academic performance are summarized. A negative correlation was found in all cases between  $n\text{-Ach}$  and the criteria of academic performance. None of the correlations reached the .05 level of significance.

The addition of the  $n\text{-Ach}$  score to the ACT composite score in a multiple regression equation served to reduce the accuracy of prediction of both criteria. The decrease in correlation was not significant at the .05 level.

The correlations of ACT composite and GPA with and without  $n\text{-Ach}$  held constant were, respectively, .377 and .388. There was no significant difference between the two correlation coefficients.

Since the respective achievement indexes  $AI_{GPA}$  and  $AI_{Psy.}$  removed

TABLE I  
*Coefficient of Correlation between Predictors and  
 Criteria of Academic Performance*  
 (N = 217)

Predictors	Criterion Variables	
	GPA	Psy.
<i>n-Ach</i>	-.023	-.028
<i>ACT</i>	.377	.422
<i>ACT + n-Ach (R)</i>	.142	.178

Note—In all cases a two-tailed test of significance was used.

The correlation between *ACT* and *n-Ach* was  $-.078$

variance accounted for by differences in aptitude, negligible  $r$ 's or  $-.063$  and  $-.050$  were obtained between each of the achievement indexes and the ACT Composite. The results indicated no prediction of over-achievement from the ACT Composite.

The problem of possible overlap of aptitude and *n-Ach* was investigated by calculating partial  $r$ 's between the ACT Composite and each of the two AI indexes with *n-Ach* held constant. Neither the correlation of  $.048$  involving the  $AI_{GPA}$  or the one of  $.053$  for the  $AI_{Psy}$  was statistically significant at the  $.05$  level.

### *Differentiation of Over- and Under-Achievement*

In view of the negative findings for *n-Ach*, it was decided to evaluate all EPPS scales as possible predictors of differences in achievement. The coefficients of correlation between the  $AI_{GPA}$  and each of the 15 variables of the EPPS varied from  $-.138$  to  $.147$ , and the coefficients between  $AI_{Psy}$  and each of the 15 scales ranged from  $-.120$  to  $.159$ . Only three of the coefficients reached the  $.05$  level of significance. Correlations of  $.147$  and  $.159$  between the Intraception Scale and  $AI_{GPA}$  and between the Intraception Scale and  $AI_{Psy}$ , respectively, were statistically significant beyond the  $.05$  level as was the correlation of  $-.138$  between the Dominance Scale and  $AI_{GPA}$ .

### *Criteria of Academic Performance*

Table I indicates that in every instance the use of psychology grades as a criterion resulted in higher coefficients of correlation than those obtained using GPA. Relative to EPPS subscales, the coefficients of correlation obtained using psychology grades as the criteria were not consistently higher than were those obtained using GPA.

*Discussion*

The results of this study offered little support for the use of the Edwards *n-Ach* scales as a supplementary predictor of academic achievement. Among several studies examined but not cited only the one by Weiss, Wertheimer, and Groesbeck (1959) presented evidence that the use of the *n-Ach* scale improved the prediction of academic achievement. It should be noted that their data were based on only 49 subjects, whereas the present sample included 217, and that different academic aptitude test scores were used in the two studies. The results are consistent with Bachman's (1964) findings, with the exception that nonsignificant and negative correlations were found in all cases between *n-Ach* and the criteria of academic performance. It should be noted that Bachman's sample included only 61 subjects, while the present sample included 217 subjects. Bachman used SAT composite scores as a measure of academic aptitude, whereas in the present study, ACT composite and subtest scores were used as measures of academic aptitude. Thus, it would seem reasonable to conclude that the Edwards *n-Ach* scale is not a useful supplement to ability test scores in the prediction of academic performance and is of little value in differentiating between over- and under-achievers.

## REFERENCES

- American College Test. *Research Service Report for Marshall University, Summary Analysis*. Iowa City, Iowa, Summer, 1965. Table LI-2.5.
- Bachman, J. G. Prediction of academic achievement using the Edwards Need Achievement Scale. *Journal of Applied Psychology*, 1964, 48, 16-19.
- Edwards, A. L. *Manual for the Edwards Personal Preference Schedule* (Rev. ed.). New York: Psychological Corporation, 1959.
- Gebhart, C. G., and Hoyt, D. T. Personality needs of under- and over-achieving freshman. *Journal of Applied Psychology*, 1958, 42, 125-128.
- Goodstein, L. D., and Heilbrum, A. B. Prediction of college achievement from the Edwards Personal Preference Schedule at three levels of intellectual ability. *Journal of Applied Psychology*, 1962, 46, 317-320.
- Krug, R. D. Over- and under-achievement and the Edwards Personal Preference Schedule. *Journal of Applied Psychology*, 1959, 43, 133-136.
- Weiss, P., Wertheimer, M., and Groesbeck, B. Achievement motivation, academic aptitude, and college grades. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1959, 19, 663-666.



## VALIDITY OF THE MMPI-168 FOR PSYCHIATRIC SCREENING<sup>1</sup>

JOHN E. OVERALL<sup>2</sup>

University of Texas Medical Branch, Galveston

JAMES N. BUTCHER

University of Minnesota, Minneapolis

SARA HUNTER

University of North Carolina, Chapel Hill

Validity of an abbreviated 168-item administration and the standard MMPI was compared with reference to discriminating psychiatric patients from normal college students. Better discrimination was obtained from clinical scale scoring than from factor scoring. The abbreviated MMPI-168 actually produced slightly better discrimination than did the longer parent instrument. Revised equations for converting MMPI-168 scores to conventional MMPI validity and clinical scale scores are presented.

THE purpose of this study was to compare the discriminant validity, for general psychiatric screening, of an abbreviated 168-item administration of the Minnesota Multiphasic Personality Inventory (MMPI) with that of the standard 373-item short form. A secondary purpose of this article was to provide new and improved equations for converting scores derived from an abbreviated administration to equivalent MMPI clinical scale scores.

Overall and Gomez-Mont (1974) have presented evidence that much of the reliable variance of the standard MMPI clinical scales is concentrated in the first 168 items. A procedure for estimating conven-

<sup>1</sup> This work was supported in part by grant DHEW 5 R01 MH14675-07.

<sup>2</sup> Requests for reprints should be addressed to John E. Overall, Psychometric Laboratory, University of Texas Medical Branch, Galveston, Texas 77550.

Copyright © 1975 by Frederic Kuder

tional clinical scale scores from raw scores obtained by applying standard MMPI scoring stencils to the first 168 items was described. Overall, Hunter, and Butcher (1973) investigated the factor structure of the first 168 items in an effort to understand further the content of the abbreviated version. Factors representing Somatization, Depression, Low Morale, Psychotic Distortion, Acting Out, plus an *Mf* Feminine Interests factor were identified, and items constituting factor scoring keys were reported. Analyses of the larger MMPI item pool were undertaken by Hunter, Overall, and Butcher (1974) to verify that the structure was not appreciably different from that of the first 168 items. As a result of these investigations, factor scoring procedures purported to represent the same basic dimensions of psychopathology were made available for the 373-item MMPI short form and for the 168-item abbreviated short form. Thus, in the present investigation, it was possible to compare the discriminant validity of the abbreviated and standard administrations both in terms of clinical scale scoring and factor scoring.

Because the writers have envisioned an important use of the MMPI to be psychiatric screening in ostensibly normal populations, the ability of the instrument to discriminate psychiatric patients from normal college students seems a reasonable basis for comparison of the validity of different scoring procedures. Although the psychiatric sample considered in this study was not matched to the college sample in age or social class, it seems appropriate to assume that the primary source of difference in MMPI clinical scale score and factor score profiles should be the degree of psychopathology.

### *Method*

A mixed clinical consisting of 431 subjects including neurotics, psychotics, personality disorders, alcoholics, and drug abusers was obtained from a state hospital, from diagnostic referrals of private patients in a university hospital, from an inpatient alcohol treatment unit, and from an outpatient drug rehabilitation unit. Males outnumbered females approximately 3 to 1 in the psychiatric sample. A normal comparison group was obtained by randomly sampling, in approximately the same sex ratio, 400 MMPI records from a larger college student sample obtained in group testing at Bethel College (Minnesota).

The MMPI item response protocols were computer scored according to four procedures: (a) standard clinical scale scoring, (b) clinical scale scoring based on the first 168 items, (c) factor scoring based on the 373-item short form, and (d) factor scoring based on the first 168

items. For each type of scoring, means and standard deviations were calculated for each variable from the normal college sample alone, and all scores were transformed to T-scores based on the college norms.

Assuming that the psychiatric sample should differ from the college sample by evidencing higher levels of psychopathology, the first analyses involved a simple determination of the number of subjects in each group who had one or more scores elevated above *T*-score of 70 according to each of the four scoring procedures. Because of the mixed sex composition of the samples, *Mf* scales or factor scores were not considered in these analyses.

Next, the method of linear discriminant function analysis was used to determine the degree of separation of the normal and psychiatric samples. Cutting points were selected to minimize classification errors in the sample data. To evaluate possible shrinkage, the discriminant function analyses were repeated with the samples split at random into primary and cross-validation samples. The discriminant function derived from the primary samples was used to assign subjects in the cross-validation samples.

### Results

The frequencies of classification based on observation of one or more elevated scores ( $T > 70$  in college population) are presented in Table 1 for factor scoring and clinical scale scoring based on 168 items and on 373 items. Because the *T*-scores were normed on the college population, the use of  $T > 70$  as a cutting point resulted in fewer misclassification of normals than of psychiatric patients. Also, because there were fewer factor scores than clinical scale scores on which elevation could occur, the proportion classified as abnormal according to factor scoring was lower than for clinical scale scoring. These differences, however, should not affect the comparisons of primary interest in this investigation. The comparability of classification based on the 168-item abbreviated short form with that obtained from the standard 373-item short form is impressive. Probably as a matter of chance in these particular data, the abbreviated MMPI-168 scoring actually yielded fewer errors of classification than did the MMPI-373 with both factor scoring and clinical scale scoring. The validity of factor scoring and clinical scale scoring appeared comparable also, when simple  $T > 70$  scale elevation was used as the criterion of abnormality. The contingency coefficients relating MMPI classification to actual group membership were essentially identical (.46, .47, .45 and .45) for the four sets of results shown in Table 1.

Next, a simple discriminant function analysis was performed on

TABLE 1  
*Frequencies of Correct and Incorrect Classification Based on One or More  
 Elevated Scales from Four Different Scoring Procedures*

*Clinical Scale Scoring 373 Items*

	College Sample	Clinical Sample
Normal	332	136
Abnormal	68	295

*Clinical Scale Scoring 168 Items*

	College Sample	Clinical Sample
Normal	331	130
Abnormal	69	301

*Factor Scoring 373 Items*

	College Sample	Clinical Sample
Normal	349	167
Abnormal	51	264

*Factor Scoring 168 Items*

	College Sample	Clinical Sample
Normal	351	167
Abnormal	49	264

each of the sets of scores.<sup>3</sup> The computer program that was used produced a percentile frequency distribution of scores on each optimally weighted composite (discriminant function) for the two groups. A cutting point was selected by inspection to minimize classification errors in the sample data. Results are presented in Table 2 for the four types of scoring.

Contrary to the results obtained from classification based on single scale elevation (Table 1), the discriminant function approach yielded considerably better results when applied to clinical scale scores than to factor scores. It is also apparent that the discriminant function approach applied to clinical scale scores did yield considerably fewer classification errors than did the criterion of single scale elevation applied to the same type of scoring. Of primary concern in this investigation, the discriminant function classification based on the 168-item ab-

<sup>3</sup> The specific discriminant function coefficients are not reported here because optimum combination of clinical scale scores depends on the psychiatric sample to be discriminated. This study was addressed to the question of which type of scoring should be expected to yield best discrimination, not the definition of a particular weighted composite for general use.

breviated short form was as accurate as the classification based on standard 373-item short form scoring.

To confirm that the superior results obtained for clinical scale scoring were not due to capitalizing on chance relationships among the larger set of clinical scale scores, as compared with factor scores, the samples were subsequently split at random into primary and cross-validation samples. Discriminant functions were fitted in Sample A for both factor scoring and clinical scale scoring, and then the accuracy of classification was assessed in Sample B. The shrinkage observed in the cross-validation samples was approximately 2 per cent. Because the cross-validation results were so similar to those cited in Table 2, their presentation in separate tables seems unnecessary.

Having verified high discriminant validity for clinical scale scores estimated from an abbreviated 168-item administration of the MMPI through use of regression transformations previously derived by Overall and Gomez-Mont (1974) from a relatively small sample of psychiatric patients, the writers calculated new and more parsimonious regression transformations from the larger combined psychiatric and college samples of this investigation. To facilitate clinical

TABLE 2  
*Frequencies of Correct and Incorrect Classification Based on Discriminant Functions from Four Different Scoring Procedures*

*Clinical Scale Scoring 373 Items*

	College Sample	Clinical Sample
Normal	342	57
Abnormal	58	374

*Clinical Scale Scoring 168 Items*

	College Sample	Clinical Sample
Normal	347	60
Abnormal	53	371

*Factor Scoring 373 Items*

	College Sample	Clinical Sample
Normal	303	118
Abnormal	97	313

*Factor Scoring 168 Items*

	College Sample	Clinical Sample
Normal	299	115
Abnormal	101	316



use, the new regression transformations were calculated to go from each individual MMPI-168 raw score to the estimated corresponding MMPI raw clinical scale score without additional scales being included in the equations. The variances of the regression transformations were also adjusted to equal the variances of the standard MMPI raw scale scores in the composite sample under consideration. Without such adjustment, the variances of least squares regression estimates are smaller than the variances of the criterion scores. The "stretching" of variances has no effect on the linear correlation between predicted and observed scores, but it is important if one is going to use existing *T*-score norms for interpretation of the clinical scale scores derived from an abbreviated administration. The regression transformations for MMPI-168 raw scores (obtained by applying standard MMPI scoring stencils to the first 168 items) are presented in Table 3. The product moment correlations between the MMPI-168 and MMPI-373 clinical scale scores observed in this sample are shown in the right-hand column.

Table 3  
*Regression Estimation of Raw Scale Scores from 168-Item MMPI*

---



---

$\hat{L} = 1.29$	$\tilde{L}$	$+ 0.31$
$\hat{F} = 1.76$	$\tilde{F}$	$+ 1.63$
$\hat{K} = 1.90$	$\tilde{K}$	$+ 2.21$
$\hat{Hs} = 1.39$	$\tilde{Hs}$	$+ 0.67$
$\hat{D} = 1.26$	$\tilde{D}$	$+ 4.60$
$\hat{Hy} = 1.37$	$\tilde{Hy}$	$+ 6.86$
$\hat{Pd} = 1.37$	$\tilde{Pd}$	$+ 5.66$
$\hat{Mf} = 1.78$	$\tilde{Mf}$	$+ 5.17$
$\hat{Pa} = 2.15$	$\tilde{Pa}$	$+ 1.50$
$\hat{Pt} = 2.39$	$\tilde{Pt}$	$+ 0.82$
$\hat{Sc} = 3.52$	$\tilde{Sc}$	$+ 3.25$
$\hat{Ma} = 1.78$	$\tilde{Ma}$	$+ 1.44$
$\hat{Si} = 3.78$	$\tilde{Si}$	$+ 1.40$

---

*Discussion*

Answers to two principal questions were sought in this investigation: Is the psychiatric screening utility of an abbreviated 168-item administration of the MMPI approximately equivalent to that of the longer conventional MMPI? Is the psychiatric screening utility of factor scoring superior to that of more familiar clinical scale scoring? The answers appear to be that the abbreviated MMPI-168 has the potential for providing just as valid general psychiatric screening results as does the considerably longer standard MMPI administration and that the factor scoring examined here does not appear to offer potential superior to clinical scale scoring for the purpose of general psychiatric screening.

Much more work must be done to evaluate the clinical utility of a shortened administration of the MMPI. At this point in time, evidence appears adequate to confirm that most of the reliable variance in the standard MMPI clinical scales is accounted for by the first 168-items (Overall and Gomez-Mont, 1974), that the content domain spanned by the first 168 items is equivalent (as judged by factor structures) to that of the larger item pool (Overall, Hunter and Butcher, 1973; Hunter, Overall and Butcher, 1974), and that scale scores derived from the first 168 items correlate quite highly with clinical scale scores derived from the larger item pool (Hedlund, Powell, and Cho, 1974; Newmark, Newmark, and Cook, 1975; Newmark and Raft, 1975). That the clinical scale scores derived from the abbreviated administration have potential equivalent to those derived from a longer standard administration for general psychiatric screening appears documented in the present study.

In the further evaluation of the MMPI-168, the authors would emphasize that the conventional longer MMPI can serve only as a limited and imperfect criterion. Comparisons should consider the longer and the abbreviated versions as alternative test instruments. Where classification results do not always agree, it cannot be assumed that the longer MMPI is always correct. The use of external criteria, as in the present investigation, appears to offer an advantage for comparison of alternative forms without assuming that either is the ultimate criterion.

## REFERENCES

- Hedlund, J. L., Powell, B. J., and Cho, D. W. The use of MMPI short forms with psychiatric patients. Paper presented at the annual American Psychological Association meeting, New Orleans, 1974.
- Hunter, S., Overall, J. E., and Butcher, J. N. Factor structure of the

- MMPI in a psychiatric population. *Multivariate Behavioral Research*, 1974, 9, 283-302.
- Newmark, C. S., Newmark, L. G., and Cook, L. The MMPI-168 with psychiatric patients. *Journal of Clinical Psychology*, 1975 (in press).
- Newmark, C. S. and Raft, D. Using an abbreviated MMPI as a screening device for medical patients. *Psychiatry in Medicine*, 1975 (in press).
- Overall, J. E. and Gomez-Mont, F. The MMPI-168 for psychiatric screening. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1974, 34, 315-319.
- Overall, J. E., Hunter, S., and Butcher, J. N. Factor structure of the MMPI-168 in a psychiatric population. *Journal of Consulting and Clinical Psychology*, 1973, 41, 284-286.

## COMPARISON OF THE STANDARD MMPI AND THE MINI-MULT IN A UNIVERSITY COUNSELING CENTER

R. B. SIMONO

University of North Carolina at Charlotte

One hundred ten undergraduates were administered the standard Minnesota Multiphasic Personality Inventory which was scored for the standard profile and for a shortened version of the same inventory. The study was designed to explore the usefulness of a short version of the MMPI in a university counseling center. Correlations were obtained between corresponding scales on both forms for males and females separately. Although Pearson product-moment correlations for both sexes were statistically significant, they were not of the magnitude to predict scale scores on one form from the other. In addition, an examination of the profiles suggested that the short version could not provide clinical data comparable to those of the standard form. Implications for further research were made.

In college or university counseling centers, the standard MMPI is often used as a diagnostic and research instrument. However, psychologists in such settings have questioned the amount of time demanded from the client in order to complete this measure and have expressed concern about the lack of interest which some clients demonstrate when asked to respond to the standard MMPI.

Kincannon (1968) took an abbreviated form of the MMPI (which he titled the "Mini-Mult") consisting of 71 items and compared it to the standard MMPI with a sample of 50 male and 50 female admissions to a psychiatric unit. Included were MMPI scales 1 to 4, 6 to 9, and L, F, and K. For the two sets of raw scores on the Mini-Mult and the MMPI, product-moment correlations ranged from .80 to .93, with the median correlation being .87. A second comparison between a similar group of 25 males and 25 females, resulted in correlations which ranged from .70 to .96 with a median of .87.

Lacks (1970) replicated the findings of Kincannon with 50 males and 44 females who were in-patients in an acute intensive treatment facility. Correlations between the comparable scales of the MMPI and the Mini-Mult ranged from .68 to .89 with a median of .83. In a study by Harford, Lubetkin and Alpert (1972) correlations between corresponding scales on the Mini-Mult and the MMPI ranged between .21 to .81 with a median  $r$  of .54. When relating the scales on the MMPI to those of the Mini-Mult, Armentrout (1970) obtained statistically significant correlations in all scale comparisons except for males on the  $F$  scale. For males  $r$ 's ranged from .09 to .73 and for females from .42 to .85. In this study a comparison of profile peaks led to the conclusion that "Mini-Mult profiles did not permit prediction of the one or two most evaluated clinical scales" on the MMPI.

The present study was designed to explore the usefulness of a short version of the MMPI in a university counseling center as well as to determine whether earlier results of investigations of the Mini-Mult could be replicated with a sample of college males and females demonstrating no gross abnormality.

### *Method*

The subjects for this study were 110 male and 45 female undergraduates who were clients in a university counseling center which offered general psychological services to a population of approximately 5,000 students. When counseling was initiated, the clients were administered the standard MMPI. The MMPI's were scored for standard profile ( $K$ -corrected) and Kincannon's scoring procedures were used to obtain Mini-Mult scores on the same set of data.

### *Results*

Table 1 shows the product-moment correlations of the comparable scales for both the standard MMPI and the Mini-Mult. The correlations between the scales for the male sample yielded values ranging from .03 to .79, with a median correlation of .60. Although 10 of the correlations were significant ( $p < .01$ ), only five scale correlations ( $K$ ,  $D$ ,  $Pa$ ,  $Pt$ ,  $Sc$ ) were high enough that one-half of the variance of one scale was accounted for by that of the other scale.

For the sample of females, correlations between scales ranged from .18 to .85 with a median correlation of .71. Although nine of the correlations were significant ( $p < .01$ ), only six of the scale correlations ( $D$ ,  $Pd$ ,  $Pa$ ,  $Pt$ ,  $Sc$ ,  $Ma$ ) were such that one-half of the variance of one scale was accounted for by that of the other scale.



TABLE I  
Correlations of the Corresponding Scale Scores for the  
Standard MMPI and the Mini-Mult

Scale	Females (N = 45)	Males (N = 110)
	<i>r</i>	<i>r</i>
<i>L</i>	.53**	.39**
<i>F</i>	.54**	.21**
<i>K</i>	.63**	.71**
<i>Hs</i>	.33	.03
<i>D</i>	.71**	.75**
<i>Hy</i>	.18	.34**
<i>Pd</i>	.81**	.60**
<i>Pa</i>	.72**	.71**
<i>Pt</i>	.79**	.79**
<i>Sc</i>	.85**	.76**
<i>Ma</i>	.77**	.59**

\*\*  $p < .01$ .

Because the MMPI is often used in a therapy setting to generate or confirm clinical hypotheses, it seemed important to explore the degree to which the Mini-Mult could predict various indices of personal/social adaptation problems in a college population. The method used, which was developed by Drake and Oetting (1959), was designed to help develop hypotheses from MMPI patterns. No scale was coded high unless it had a *T*-score of 55 or above. Only the three highest scales were selected to represent the high coding of the profile. After the three highest scales were chosen, the numbers of the scales were arranged in numerical order regardless of the relative size of the *T*-scores. This procedure was followed for both the standard MMPI and the Mini-Mult in order to obtain a comparison of profile peaks on the two comparable forms of this personality inventory.

After looking at the profile pairs for females according to the Drake and Oetting procedures and noting only the highest three scales regardless of relative size of *T*-scores, the investigator found that 15 individuals had three scales in the same order on both forms, seven had two scales in the same order in each profile, and 14 had one scale which showed up high in one profile and also in the other. Nine female profile pairs showed no similarity in profile peaks through using the Drake and Oetting procedures. The 110 male profile pairs were then examined also through employing the Drake and Oetting procedures. It was found that seven pairs had three scales in the same order on each form, 45 pairs had two scales in the same order in each form and 35 pairs had one identical scale in each form. Twenty-three pairs of profiles showed no relationship to each other.

*Discussion*

The results of the present investigation of the comparability of the standard Minnesota Multiphasic Personality Inventory and the Mini-Mult were dissimilar from those found by Kincannon (1968) and Lacks (1970) but, similar to those obtained in the comparability studies of Harford et al. (1972) and Armentrout (1970).

Although there were significant correlations found between the scales of the MMPI and corresponding scales on the Mini-Mult, the results reinforced doubts that the Mini-Mult could be used in a University counseling center to provide clinical data comparable to those available from the standard MMPI. Neither the present study nor Armentrout's (1970) work approached the level of comparability seen in the work of Kincannon (1968) and Lacks (1970). Reinforced was the suggestion made by Harford et al. (1972) that the comparability of the two forms is enhanced when working with a "more pathologically severe population."

Although the Mini-Mult did not provide profile peaks equivalent to those in the standard version of the MMPI, it did furnish personality data which met criteria used to indicate emotional tension or difficulty in personal adaptation as measured by the MMPI. Because measures such as the Mini-Mult offer brief methods of personality assessment, additional research as well as clinical observation might determine the validity of the Mini-Mult as a separate personality measure.

## REFERENCES

- Armentrout, J. A. Correspondence of the MMPI and Mini-Mult in a college population. *Journal of Clinical Psychology*, 1970, 26, 493-495.
- Drake, L. E. and Oetting, E. R. *An MMPI code book for counselors*. Minneapolis: University of Minnesota Press, 1959.
- Harford, T., Lubetkin, B., and Alpert, G. Comparison of the standard MMPI and the Mini-Mult in a psychiatric out-patient clinic. *Journal of Consulting and Clinical Psychology*, 1972, 39, 243-245.
- Kincannon, J. C. Prediction of the standard MMPI scale scores from 71 items; The Mini-Mult. *Journal of Consulting and Clinical Psychology*, 1968, 32, 319-325.
- Lacks, P. B. Further investigation of the Mini-Mult. *Journal of Consulting and Clinical Psychology*, 1970, 35, 126-127.

THE FACTORIAL VALIDITY OF  
THE PIERS-HARRIS CHILDREN'S SELF-CONCEPT  
SCALE FOR EACH OF THREE SAMPLES OF  
ELEMENTARY, JUNIOR HIGH, AND SENIOR HIGH  
SCHOOL STUDENTS IN A LARGE  
METROPOLITAN SCHOOL DISTRICT

WILLIAM B. MICHAEL AND ROBERT A. SMITH  
University of Southern California

JOAN J. MICHAEL  
California State University, Long Beach

For each of three samples of 299 elementary school pupils, 302 junior high school pupils, and 300 senior high school students in a large metropolitan school district, factor analyses of the intercorrelations of the responses to 80 items in The Piers-Harris Children's Self-Concept Scale yielded three major dimensions that were essentially invariant across the three samples: (a) physical appearance, (b) socially unacceptable (bad) behavior, and (c) academic or school status. Variance within a complex domain of emotionality was differentiated among a number of factors such as anxiety, abasement, self-contentment, and self-dissatisfaction that were not invariant across samples. In both the junior high school and senior high school samples identifiable factors of popularity and perceived psychomotor coordination appeared. Implications for writing of items to represent the constructs associated with self-concept are discussed.

AMONG several self-concept measures available to researchers and school personnel The Piers-Harris Children's Self-Concept Scale (Piers and Harris, 1969) was evaluated by Shreve (1973) to show the greatest promise according to criteria posed in the Technical Standards for Educational and Psychological Tests (French and Michael, 1966). In their manual Piers and Harris (1969) reported a factor analy-

sis of the scale of 80 yes-no items that had been administered to a sample of 457 sixth grade pupils. Six factors were tentatively identified as *"undesirable or bad" behavior; intellectual and school status, physical appearance and attributes, anxiety, popularity, and happiness and satisfaction.*

The two-fold purpose of this investigation was to determine for each of three samples of 299 elementary school pupils, 302 junior high school students, and 300 senior high school students in one of the nation's largest metropolitan school districts the factorial dimensions of The Piers-Harris Children's Self-Concept Scale and to ascertain whether the constructs associated with the six factors reported by Piers and Harris could be replicated. Determination of the factorial structure of this scale could enhance its utility in the identification of the measurable constructs underlying the complex entity called self-concept, could provide an improved basis for use of the instrument in diagnostic and counseling activities, and could suggest possible revisions in or additions to items in the scale.

### *Methodology*

Intercorrelations (phi coefficients) of the responses to the 80 items were found for each of the three samples, and in each sample a varimax factor rotation was undertaken of all principal components with eigenvalues in excess of unity (Dixon, 1969).

In general, with one or two exceptions, a factor was identified and is cited in this paper whenever it yielded a loading of at least .60 on one item, of at least .50 on a second item, and of at least .30 on each of two or more other items.

### *Findings*

The results of the investigation are reported in terms of (a) the citation of three major factorial dimensions that were essentially replicated across the three samples studied in three of the key dimensions reported by Piers (1969), (b) a description of the factorial structure of the broad but central domain of emotionality, and (c) a brief exposition of other factors of secondary significance or concern.

#### *Factors More or Less Invariant across Samples*

Across each of the three educational levels the three following factors appeared to be relatively invariant: physical appearance, socially unacceptable (bad) behavior, and academic competence reflecting

school and intellectual status. For each of the groups representing a different educational level, a single well defined factor regarding physical characteristics emerged. Relative to the construct revealing socially undesirable behavior only one factor appeared for the elementary and junior high school groups, but for the senior high school sample two factors resulted—the second one reflecting to a large extent home difficulties. In the instance of the academic dimension, however, one factor indicating mostly observable activities or behaviors in the sets of items and a second factor involving sets of items revealing primarily passive or reported subjective feelings appeared at each of the three educational levels. The factor analytic results indicating loadings on individual test items at least equal to .30 for the physical appearance dimension may be summarized for the elementary school (ES), junior high school (JHS), and senior high school (SHS) samples as follows:

	<i>Factor I—Physical Appearance</i>	1 (ES)	1 (JHS)	1 (SHS)
54	I am good looking.	.75	.49	.75
60	I have a pleasant face.	.71	.43	—
73	I have a good figure.	.68	—	.40
41	I have nice hair.	.67	.70	.66
69	I am popular with girls.	.58	.41	.45
29	I have pretty eyes.	.58	—	—
15	I am strong.	—	.58	—
57	I am popular with boys.	—	.57	.52
21	I am good in my school work.	.54	—	—
5	I am smart.	.50	—	.35
8	My looks bother me.	—	—	-.46
49	My classmates in school think			
	I have good ideas.	.40	—	—
33	My friends like my ideas.	.39	.36	—
36	I am lucky.	—	.37	—
70	I am a good reader.	.35	—	—
27	I am an important member of			
	my class.	.35	—	—
52	I am cheerful.	.34	—	—
80	I am a good person.	.34	—	—
9	When I grow up, I will be an			
	important person.	.34	—	—

For the dimension of socially unacceptable, or so-called bad, be-



havior, which included the appearance of two factors at the senior high school level, the following outcomes were observed for each of the three school groups:

<i>Factor II—Socially Unacceptable ("Bad") Behavior</i>		II (ES)	II (JHS)	IIA (SHS)	IIB (SHS)
22	I do many bad things.	.69	.39	.54	—
34	I often get into trouble.	.50	—	.65	—
48	I am often mean to other people.	.64	.57	.49	—
56	I get into a lot of fights.	.54	.66	.57	—
12	I am well behaved in school.	-.44	—	-.48	—
78	I think bad thoughts.	.42	.45	.33	—
14	I cause trouble to my family.	.38	.33	—	—
68	I lose my temper easily.	.37	—	—	—
13	It is usually my fault when something goes wrong.	—	—	.42	—
38	My parents expect too much of me.	—	.41	—	—
25	I behave badly at home.	.37	.36	—	.64
62	I am picked on at home.	—	.51	—	.67
59	My family is disappointed in me.	—	.48	—	.57
44	I sleep well at night.	—	—	—	-.36
61	When I try to make something, everything seems to go wrong.	—	—	—	.36

For what appeared to be primarily activity-oriented items associated with academic competence, the factor analytic results relative to each of the three samples at differing educational levels were as follows:

<i>Factor IIIA—Academic or School Status Embodying Many Activity-oriented Items</i>		IIIA (ES)	IIIA (JHS)	IIIA (SHS)
42	I often volunteer in school.	.65	—	.58
36	I can give a good report in front of the class.	.53	.61	.66

49	My classmates in school think I have good ideas.	.53	.34	—
16	I have good ideas.	.37	—	—
33	My friends like my ideas.	.34	—	—
7	I get nervous when the teacher calls on me.	—	—	-.42
12	I am well behaved in school.	—	.39	—
75	I am always dropping or breaking things.	—	-.35	—
63	I am a leader in games and sports.	.31	—	—
36	I am lucky.	—	.34	—
66	I forget what I learn.	—	—	-.38
6	I am shy.	—	—	-.34

In the instance of items that tended to reflect somewhat passive and subjective states of feeling regarding school competence in the absence of much overt activity, the factor analytic data relative to each of the same three groups representing differing educational levels were as follows:

<i>Factor IIIB—Academic or School Status Reflecting Subjective Feelings Rather Than Overt Activities</i>		IIIB (ES)	IIIB (JHS)	IIIB (SHS)
26	I am slow in finishing my school work.	-.61	-.31	-.67
		.59	—	.51
70	I am a good reader.	—	.42	.57
21	I am good in my school work.	—	.79	—
5	I am smart.	—	—	—
61	When I try to make something, everything seems to go wrong.	—	—	-.32
38	My parents expect too much of me.	-.45	—	—
		-.35	—	—
22	I do many bad things.	-.35	—	—
66	I forget what I learn.	—	—	—

#### *Factors Pertaining to Components of Emotionality*

Relative to the factors that arose from clusters of mainly nonbehaviorally stated items dealing with such components of emotionality as

anxiety, self-depreciation (abasement), self-actualization (self-satisfaction or happiness), and status or power needs the factorial picture was not only somewhat ambiguous or ill-defined, within each of the three samples at varying school levels, but also quite variable from one school level to another.

For the elementary school group, two rotated factors (I and V)—the first one being made up of more behaviorally stated items than was the second one—appeared to be sufficiently well described to permit respective interpretations of self-depreciation and of anxiety embodying withdrawal and alienation tendencies (although the patterns of three or four item loadings on essentially a residual factor not to be cited also suggested the possible presence of alienation accompanied by ineptness, hypochondriasis, and egocentrism):

*Factor IVA (ES)—Self-Depreciation (Abasement)*

1 My classmates make fun of me.	.70
75 I am always dropping or breaking things.	.68
79 I cry easily.	.67
61 When I try to make something, everything seems to go wrong.	.63
53 I am dumb about most things.	.47
59 My family is disappointed in me.	.36
13 It is usually my fault when something goes wrong.	.34
11 I am unpopular.	.31

*Factor IVB (ES)—Anxiety Involving Withdrawing Behavior*

28 I am nervous.	.68
10 I get worried when we have tests in school.	.59
6 I am shy.	.57
37 I worry a lot.	.56
20 I give up easily.	.52
43 I wish I were different.	.38
7 I get nervous when the teacher calls on me.	.37
50 I am unhappy.	.30

At the junior high school level two separate identifiable abasement-oriented factors (I and VII), which might have been anticipated to fuse as one general factor, did emerge as did two contrasting factors (V and XVIII) of self-contentment (happiness) and self-dissatisfaction reflecting anxiety and alienation characteristics.

*Factor IVA (JHS)—Abasement or Self-Depreciation*

20 I give up easily.	.66
11 I am unpopular.	.59
48 I am often mean to other people.	.50
8 My looks bother me.	.45
6 I am shy.	.42
79 I cry easily.	.30

*Factor IVB (JHS)—Self-Depreciation Involving Alienation and Self-Pity*

61 When I try to make something, everything seems to go wrong.	.66
1 My classmates make fun of me.	.63
13 It is usually my fault when something goes wrong.	.43
45 I hate school.	.41
64 I am clumsy.	.38
65 In games and sports, I watch instead of play.	.37
62 I am picked on at home.	.32

*Factor IVC (JHS)—Self-Contentment (Happiness)*

52 I am cheerful.	.67
80 I am a good person.	.62
2 I am a happy person.	.58
12 I am well behaved in school.	.42
35 I am obedient at home.	.38
60 I have a pleasant face.	.36

*Factor IVD (JHS)—Self-Dissatisfaction (Unhappiness) Involving Anxiety and Instability*

43 I wish I were different.	.77
40 I feel left out of things.	.60
39 I like being the way I am.	-.58
4 I am often sad.	.53
51 I have many friends.	-.42
28 I am nervous.	.42
74 I am often afraid.	.32
8 My looks bother me.	.32

Two other factors each consisting of three item variables suggested two additional dimensions of alienation and self-centeredness,

although a dependable identification was not possible. These results are not reported.

At the senior high school level only one clearly identifiable factor in the area of emotionality was found. Relative to self-dissatisfaction or unhappiness this factor was described in terms of the following items and their loadings:

*Factor IV (SHS)—Self-Dissatisfaction Reflecting Alienation and Anxiety*

58 People pick on me.	.71
40 I feel left out of things.	.51
1 My classmates make fun of me.	.44
6 I am shy.	.42
7 I get nervous when the teacher calls on me.	.39
50 I am unhappy.	.35

It should be mentioned that there was for the sample of senior high school students the strong suggestion of a status-power factor as indicated by loadings on four items:

63 I am a leader in games and sports.	.69
27 I am an important member in my class.	.64
15 I am strong.	.39
36 I am lucky.	.37

This perception of one's having status and influence might reflect the development of a need for power in the social order on the part of adolescents in senior high school who have been acquiring an appreciation of the importance of power in adolescent and adult cultures.

*Other Factors*

Although not identifiable for all three student groups, two interpretable factors did emerge that were common to the junior high school and senior high school samples: perceived popularity and perceived psychomotor coordination (skill). Thus for the factor of popularity the following results were obtained:

*Factor V—Popularity*

	V (JHS)	V (SHS)
51 I have many friends.	.65	.60
3 It is hard for me to make friends.	-.44	-.73



69	I am popular with girls.	.56	.32
28	I am nervous.	—	— .47
6	I am shy.	— .34	—
58	People pick on me.	— .33	—

The factor involving perception of psychomotor coordination was relatively clearly defined by two items in particular as follows:

<i>Factor VI—(Perceived) Psychomotor Coordination</i>		VI (JHS)	VI (SHS)
19	I am good at making things with my hands.	.74	.79
23	I can draw well.	.72	.74
16	I have good ideas.	.40	.46
33	My friends like my ideas.	.36	—

### *Discussion*

The identification in each of the three sample differing in educational level of the same three constructs of *physical appearance*, *socially unacceptable* ("bad") *behavior*, and *academic or school competence* (which was portrayed by two factors) replicated the first three constructs in the factor analytic investigation reported by Piers (1969). In what the writers preferred to call a domain of emotionality the results were somewhat complex. At the elementary school level two factors of *self-depreciation* (*abasement*) and *anxiety* were evident, whereas at the junior high school level four factors of *abasement or self-depreciation*, *self-depreciation* (with overtones of *alienation*, *self-pity*, and *masochism*), *self-contentment* (*happiness*), and *self-dissatisfaction* (*unhappiness*) involving *anxiety feelings* and *instability* emerged. Yet, at the senior high school level only one readily identifiable factor of *self-dissatisfaction* (*unhappiness*) reflecting *alienation* and *anxiety* components appeared. Thus the factors of *anxiety* and *happiness* reported by Piers were only partially replicated, and in the instance of the junior high school sample two *abasement* factors, one factor reflecting *unhappiness*, and one factor suggesting a positive affect of *happiness* were seemingly required to cover the measurable variance in emotionality. For senior high school students the one dimension of negative affect that could be isolated appeared to cut across the two dimensions reported by Piers (1969). This dimension actively constituted a reflection (negative direction) of the *happiness* factor described by Piers and incorporated elements of the *anxiety* factor. In the junior high school and in the senior high school, but not in the elementary school sample, the factor interpreted to be *popularity* was also replicated. In addition, the suggestion of a new factor involv-

ing the perception of competence in psychomotor skills for the junior high school and senior high school samples was apparent, and the hint of a status or power factor was evident for the sample of senior high school students.

Although the conclusion could be made that several of the factorial dimensions identified by Piers were replicated in the samples studied, the complex domain of emotionality involving such constructs as happiness (self-satisfaction or self-actualization), unhappiness (self-dissatisfaction or lack of self-actualization), and anxiety with overtones of abasement, self-depreciation, masochism, and guilt yielded results that were not too nearly congruent with the factor analytic findings cited by Piers who used essentially the same factor analytic procedures as those employed by the writers. Even though methodological difficulties in use of principal components extraction and in the subsequent varimax rotation might account in part for the inability of the investigation to yield as psychologically meaningful results as might be possible, there was the strong suggestion that the items making up the domain of emotionality were open to a greater variety of interpretations and to a more subjective evaluation upon being read than were those items associated with physical appearance, intellectual or academic status, and so-called bad or troublesome behaviors. Thus it would appear that major efforts in item writing to operationalize affective constructs that can be anchored to as comprehensive and clear-cut theoretical formulation of the nature of affective behaviors in the self-concept as is feasible would enhance the factorial validity as well as the utility of a revised form of The Piers-Harris Children's Self-Concept Scale.

## REFERENCES

- Dixon, W. J. (Ed.). BMDX72 factor analysis. *University of California Publications in automated computation, No. 3, BMD, Biomedical Computer Programs, X series supplement*. Berkeley and Los Angeles: University of California Press, 1969.
- French, J. W. and Michael, W. B. (Eds.). *Standards for educational and psychological tests and manuals*. Washington, D. C.: American Psychological Associates, 1966.
- Piers, E. V. Manual for the Piers-Harris Children's Self-Concept Scale. Nashville, Tenn.: Counselor Recordings and Tests, Box 6184, Acklen Station, 1969.
- Piers, E. V. and Harris, D. B. The Piers-Harris Children's Self-Concept Scale. Nashville, Tenn.: Counselor Recordings and Tests, Box 6184, Acklen Station, 1969.
- Shreve, E. E. A critical analysis and evaluation of evidence regarding the reliability and validity of four selected measures of self-concept. Unpublished doctoral dissertation, University of Southern California, 1973.

## THE CONTENT AND CONSTRUCT VALIDITY OF THE BARTH SCALE: ASSUMPTIONS OF OPEN EDUCATION<sup>1</sup>

ANTHONY J. COLETTA

William Paterson College of New Jersey

ROBERT K. GABLE

University of Connecticut

Latent partition analysis was used to generate content categories from judgmental data gathered on Barth Scale items from 23 open education experts. Principal component analysis was employed to examine constructs derived from response data gathered from 78 open and 113 traditional teachers. Alpha internal consistency reliabilities were developed for item clusters defining each dimension. Relationships between the judgmental categories and response dimensions facilitated naming the derived dimensions. Evidence of content and construct validity and of internal consistency reliability indicate appropriate scoring and interpretation for Barth Scale items.

ALTHOUGH an increasing number of school systems have recently adopted open education practices, the approach has been subjected to little systematic research. In particular need of study is the development of instrumentation for research in the area of open education. Bussis and Chittenden (1970) have cited the importance of the Barth Scale for describing how teachers view their role and the process of children's learning. The Barth Scale could prove useful to the school

---

<sup>1</sup> The authors gratefully acknowledge the assistance of the content experts and teachers who participated in this study, especially: Terry Denny, Beatrice Gross, David Hawkins, George Hein, John Holt, Vito Perrone, Joseph Randazzo, Charles Rathbone, Vincent Rogers, Bernard Spodek, and Lillian Weber. The support of the University of Connecticut Computer Center under the National Science Foundation Grant GJ-9 is acknowledged.

system as well as individual teachers for examining beliefs regarding learning, knowledge, and evaluation during the formative stages of implementing open classrooms.

The Barth Scale contains 28 Likert items generated by Roland S. Barth (1970, 1971) to examine the many written and verbal statements offered by open educators.<sup>2</sup> Essentially, he endeavored to make explicit many of the assumptions which underlie the practices of open education.

The primary purpose of this paper was to report an examination of the content and construct validity of the Barth Scale. Latent participation analysis (Wiley, 1967; Gable and Pruzek, 1972) was employed to generate item content categories for judgmental data gathered from open education experts; factor analysis was used to describe constructs generated from response data gathered from open and traditional elementary school teachers. A secondary purpose of this paper was to examine the relationships between the judgmental categories (content validity) and the response dimensions (construct validity). The results of these analyses will be presented separately.

### *Method*

*Study I: Content Validity.* Latent partition analysis (LPA) was employed to study the item universe sampled by the Barth Scale items. The specific purposes of this content validity study included: (a) to examine judgmental data for meaningful content categories which reflect judges' sortings of items into mutually exclusive content piles, (b) to explore the association between categories for the possible merging of categories, and (c) to display resultant categories which were generated from the judgmental data.

### *Sample*

The sample employed for this aspect of the study consisted of 44 American open education experts identified from a group of open education authorities listed in the study by Walberg and Thomas (1971). Twenty-three judges responded; they included writers, professors, practitioners, supervisors, and consultants, experienced in teaching or observing in open classrooms.

<sup>2</sup> The version of the *Barth Scale* reported in *Phi Delta Kappan*, October 1971, contained 29 items. At Barth's suggestion, items 10 and 22 were collapsed into one item in this study. For ease of interpretation though, the original Barth item numbers will be used in the tables in this paper.

*Procedure*

The Barth Scale items were typed on individual slips of paper and mailed to the experts along with a brief letter of instructions asking the experts to sort the statements into mutually exclusive categories on the basis of similar item content. The judges' categories, called manifest categories, were submitted to latent partition analysis. The LPA procedure results in a joint proportion matrix (of order 28 by 28) where each entry indexes the proportion of sorters who placed a given pair of items in the same manifest category. From this matrix, a latent category matrix is derived; entries within this matrix index for each item the derived (latent) category to which the item belongs (Gable and Pruzek, 1972; Wiley, 1967).

*Results and Discussion*

The latent category matrix is presented in Table 1. An examination

TABLE 1  
*Derived Approximation to Latent Category Matrix for 28 Items with 7 Categories*

Item Number	CATEGORY NUMBER						
	1	2	3	4	5	6	7
1	112	0	1	-8	-6	2	0
2	112	-1	2	7	-13	-6	-8
3	98	0	0	-15	0	9	6
6	48	3	9	18	27	0	4
18	2	43	-4	42	21	0	-20
20	-6	108	4	-15	1	0	9
21	-2	109	-5	3	-1	0	-2
22*	15	93	-6	6	10	-11	-1
23	-5	115	-4	-3	-4	0	2
24	1	108	-4	-11	-10	5	6
25	-6	24	70	12	10	0	-14
26	4	-11	104	-6	0	-3	9
27	-2	-12	102	17	-2	7	-11
28	4	0	108	-15	1	-8	12
29	-3	47	55	12	-10	9	-11
13	14	-4	-7	67	-13	-4	35
16	-6	-8	-4	114	2	2	7
17	-5	-5	4	117	-7	1	-3
19	10	23	-4	59	18	2	-13
4	-6	-2	-2	19	114	-4	-12
5	50	1	-5	10	54	-11	3
7	-30	-4	11	-15	112	1	7
8	6	0	-5	0	106	-4	0
9	13	-1	-3	-22	20	17	11
11	8	-5	-2	1	5	92	-3
12	-5	4	0	0	-3	104	2
14	-6	2	3	5	2	4	89
15	1	4	1	0	0	-3	93

Note.—Rows were reordered to facilitate interpretations; all entries have been multiplied by 100

\* Items 22 and 10 from the original Barth Scale reported in *Phi Delta Kappan*, October, 1971 were combined into one item in this study.



of the entries in Table 1 denotes that seven latent categories were obtained. Twenty-six items with high loadings on only one category were selected for naming the categories. The following is a description of each category in terms of item content.

Category 1 was called Exploration (EX), since the item content described the child's natural inclination to explore when learning. The second category was called Evaluation (EV) with each item describing the philosophy of assessment in the open classroom. Category 3 was titled Knowledge (K); all items in this cluster deal with the open educator's views concerning the role of knowledge. Categories 4 and 7 were labeled Intellectual Development (ID), since the items referred to the process of cognitive development. Category 5 was called Choice (C); items defining this grouping suggested the importance of allowing children a choice in learning. It should be noted that item 5 contributed slightly to the naming of this category but also loaded moderately on Category 1: Exploration. Apparently the experts perceived the phrase "active exploration in a rich environment" as denoting both exploration and choice. The name given to Category 6 was Involvement (I) as both statements defining the category were concerned with the needs for children to share their learning experiences with others.

Although the latent category matrix revealed several clearly defined subuniverses of item content for the judgmental data, two categories reflected similar subuniverses. Examination of Table 2, containing the indices of association between the pairs of latent categories, supported the similarity of the item content in Categories 4 and 7. Thus, these two item clusters labeled Intellectual Development were combined.

In summary, the LPA method provided an objective means of

TABLE 2  
*Indices of Association between Latent Categories*

CATEGORY NUMBER							
	1	2	3	4	5	6	7
1							
2	80						
3	5	70					
4	0	11	72				
5	22	11	12	64			
6	33	5	9	15	50		
7	31	5	4	9	25	80	
	20	0	15	45	22	22	109

Note.—Entries in this matrix lie between zero and unity when the model fits the data. If the matrix is essentially diagonal, most manifest categories can be said to result from differential splitting of the same latent categories. Diagonal entries estimate the probability that any two items in that category will, in a new partition, be sorted into the same manifest category; off-diagonal entries estimate the probability that two items from two different latent categories will be placed in the same manifest category. All decimal points have been omitted.

studying content validity by identifying subuniverses of item content. The identification of such content categories based on judgmental data gathered from content experts contributes greatly to the conceptual understanding of the constructs generated from a factor analysis of item response data. The results of such a factor analysis of response data are presented in the next section.

### *Method*

*Study II: Construct Validity.* Factor analysis was employed to examine construct validity by identifying the underlying dimensions or constructs which explain item response interrelationships.

### *Sample*

The sample for the construct validity study consisted of 191 elementary school teachers from the eastern United States. Seventy-eight open and 113 traditional teachers from city, rural, and suburban schools participated from the following states or districts: Florida, Washington, D. C., Maryland, New Jersey, New York, Connecticut, Massachusetts, and New Hampshire.

### *Procedure*

Since the items on the Barth Scale tend to be grouped by Barth on the basis of similar item content, the items were randomly recorded before administering them to the teachers. For ease of interpretation, though, the original item numbers will be used in the tables presented in this section.

Each teacher responded to 28 items on a 5-point Likert scale ranging from strongly agree to strongly disagree. A 28 by 28 item intercorrelation matrix was generated and submitted to a principal components analysis followed by an obliquimax transformation (Hofmann, 1970). The derived dimensions or constructs described relationships between the Barth Scale items for actual response data.

In the section which follows, primary emphasis is placed on naming and interpreting the constructs, while attention is also given to the relationships between the final judgmental categories generated by the latent partition analysis and the response data constructs.

### *Results and Discussion*

Through using the standard root criterion of 1.0, eight components were derived; seven of these components were defined by at least two

items with loadings above .35. The component loading matrix is presented in Table 3.

Table 4 contains the factor names, original LPA item codes, Barth Scale item stems, and factor loadings. The naming of the derived factors was facilitated greatly by considering the derived content categories as noted by the LPA item codes.

Factor I was called *Curricular Flexibility*. Items defining the dimensions reflect the questioning of the existence of a minimum body of knowledge and the children's right and competence in making decisions concerning what they are to learn. Teachers tending to agree with the item content defining this factor would appear to be flexible in deciding what children will learn. These same teachers seem to consider the acquisition of knowledge a shared responsibility between themselves and the child who is the agent of his own learning and who

TABLE 3  
*Component Loading Matrix Using an Obliquimax Transformation*

Item	Components						
	I	II	III	IV	V	VI	VII
28	76						
7	50						
8	36						
23							42
29		59		36			
15		59					
17		53		44	53		
14		50					42
9		45					
20		40					47
19		35					
25			65				
27			58				
20			41			46	
22			36				
11				70			
12				56			
2				52			
4				43			
5					79		
26					50		
18					43		
24						71	
21						63	
1						57	
3						38	50
6							65
16							57
							36

Note.—All entries multiplied by 100. Only entries  $\geq 35$  were included, rows were recorded for ease of interpretation.

TABLE 4

*Factor Names, Original LPA Item Codes, Barth Scale Item Stems and Component Loadings for Derived Component Solution*

Factor	Original LPA Code and Item Number		Item Summary	Loading
<b>I</b>				
Curricular Flexibility	K	28	questionable minimum body of knowledge	.76
	C	7	right to make decisions	.50
	C	8	choice in selection of materials	.36
<b>II</b>				
Intellectual Development	K	29	knowledge resides in the knower	.59
	Ev	23	observe over a long period of time	.59
	ID	15	similar stages of intellectual development	.53
	ID	17	abstractions follow experiences	.50
	ID	14	learn at own rate and style	.45
	C	9	engage in high interest activities	.40
	Ev	20	measured qualities not important	.35
<b>III</b>				
Evaluating the Child	ID	19	errors expected and desired	.65
	K	25	qualities of being are more important	.58
	K	27	knowledge is personal	.41
	Ev	20	measured qualities not important	.36
<b>IV</b>				
Learning Through Involvement	Ev	22, 10*	involvement-learning takes place—best assessed by direct observation	.70 .56
	I	11	collaborate in exploring	.52
	I	12	share something important	.44
	ID	15	similar stages of intellectual growth	.43
	Ex	2	self-perpetuating exploratory behavior	.36
	K	29	knowledge resides in knower	
<b>V</b>				
Learning Facilitators	C	4	confidence needed for learning and choices	.79
	C	5	exploration in rich environment helps learning	.50
	K	26	knowledge is personal integration of experience	.43
<b>VI</b>				
Evaluating the Child's Work	Ev	18	verification of materials	.71
	Ev	24	work is best measure of work	.63
	Ev	21	negative effect of objective measures	.57
	K	27	knowledge is personal	.46
	Ex	1	exploration independent of adults	.38
<b>VII</b>				
Learning Through Exploration	Ex	3	exploratory behavior if not threatened	.65
	Ex	6	play-predominant mode of learning	.57
	Ex	1	exploration independent of adults	.50
	C	9	engage in high interest activities	.47
	ID	17	abstractions follow experience	.42
	C	8	choice in selection of materials	.40
	ID	16	concrete follows abstract	.36

\* As suggested by Barth, items 22 and 10 from the original Barth Scale reported in *Phi Delta Kappan*, October, 1971 were combined into one item in this study.

has the competence to make decisions pertaining to what he will learn. Since teachers are faced with the day-to-day problems of teaching content and of making or encouraging curricular decisions, it is reasonable that a dimension named Curricular Flexibility would emerge from the response data. Thus, high scores on Factor I would be obtained by teachers with positive attitudes towards curricular flexibility.

Additional support is found for this interpretation of Factor I by examining the original LPA item codes in Table 4. Factor I was defined by one item judged to reflect knowledge (Category 3) in the categorical data study; and two items were judged to reflect choice (Category 5). The remaining factor descriptions were generated after a similar consideration of the derived LPA item content categories.

Factor II was called *Intellectual Development* as it is defined mainly by items judged to be descriptive of the child's intellectual development (See Original LPS codes in Table 4). Supportive of the items depicting intellectual development are statements concerned with the resulting appropriate evaluation philosophies.

Agreement with the items in Factor II would show the teacher's understanding that children's intellectual development is not always determined by verbal responses and that there is a need for an extended time period to assess the effects of the school experience on each child. Further, teachers in accord with the items tend to believe that children pass through similar stages of intellectual growth, along a sequence from concrete experience to verbal abstractions, at their own rate and in their own style. Thus, a teacher's score on the items in Factor II would indicate his beliefs as to how children develop intellectually.

Factor III was defined by items concerning the teacher's judgment of the child's personal qualities rather than his work and, therefore, was titled *Evaluating the Child*. Teachers in agreement with the items apparently indicated the conviction that although difficult to measure, qualities needed in the search for knowledge (motivation, independence, and perserverance) are more important than those qualities which are amenable to measurement.

Whereas the experts in the content validity study tended to sort all the items dealing with evaluation into one category (II, Evaluation), the response data dimensions suggested a more specific interpretation. The teachers apparently perceived evaluations as a twofold process, separating the child as a person from the child's work. That is, the response data indicated that the items would be highly significant within the framework of two aspects of evaluation reflected in Factor II and Factor VI, called *Evaluating the Child's Work*. Ostensibly, the



personal nature of the teacher-child relationship may explain the teachers' interpretation of the items.

Factor IV was called *Learning Through Involvement*, as it emphasizes child involvement as a key issue in learning. Such involvement is supported by collaboration in exploring and sharing with others. Items contributing less to the naming of the factor state that children pass through similar stages of intellectual growth and that exploratory behavior is self-perpetuating. These items partially contribute to the naming of the factor since children in similar stages are more likely to become involved in exploration. Such investigation encourages new discoveries leading to further exploratory behavior. Teachers who agree with these items apparently believe that involvement with an activity stimulates learning which the child may want to communicate to other interested children.

Factor V was called *Learning Facilities*, as items state the need for confidence, active exploration, and personal integration of experiences as facilitators in the learning process. Teachers scoring highly on these items tend to believe that confidence assists a child in making responsible choices regarding his own learning. Further, they tend to believe that the classroom environment must contain rich and varied manipulative materials designed to facilitate learning. Finally, teachers scoring highly on this dimension indicate that learning is facilitated when knowledge is personally integrated from experience and is not cut into separate disciplines. Therefore, a teacher who earns a score on Factor V would be indicating the extent to which he believes that confidence, active exploration, and personal integration of experience facilitate learning.

Factor VI was named *Evaluating the Child's Work*. The items defining this factor emphasize the importance of the child's materials and actual work in evaluating his performance. Teachers in agreement with the items would seem to perceive materials as providing feedback information to the child which verifies whether he has answered the question or solved the problem. The same teachers tend to place much emphasis on their intuitive judgment of the child's work rather than on objective tests. They no doubt would maintain that objective measures create needless stress which results in the negative effect on learning for the young child. Moreover, those agreeing with the items tend to believe that systematic evaluation fails to measure accurately the personal and unique knowledge which a child possesses. Overall, a score on Factor VI exhibits the teacher's beliefs regarding the importance of materials and intuitive judgment rather than objective measures as the focal points in evaluation.

Factor VII was called *Learning Through Exploration*. The construct

is concerned with the exploratory nature of learning. Teachers scoring highly on this construct probably believe that if unthreatened the child will display exploratory behavior independent of adults. Teachers in agreement with the items also appear to maintain that play, the most natural kind of exploration, is not distinguished from work in early childhood; and that when the child is given considerable choice, he is likely to choose activities of high interest. The construct was defined by items clearly judged to reflect the role of exploration in learning. (See original LPA code in Table 4) Thus, a respondent's score on Factor VII indicates his convictions regarding the importance of exploration in learning.

### *Factor Intercorrelations*

The intercorrelations of the primary axes were generated. The magnitudes of the correlations (range =  $-27$  to  $+.23$ ;  $r = +.16$ ) did not suggest the need for collapsing the factors into a fewer number of dimensions.

### *Reliability*

The alpha internal consistency reliabilities of the derived factors were estimated by calculating the average of all possible combinations of item correlations and employing the Spearman-Brown formula (Stanley, 1957). The factor names, number of items per factor, and the resulting reliabilities were as follows: Curricular Flexibility—3, .64; Intellectual Development—7, .74; Evaluating the Child—4, .73; Learning Through Involvement—6, .73; Learning Facilitators—3, .62; Evaluating the Child's Work—5, .72; Learning Through Exploration—7, .76. Examination of the reliabilities indicates that several dimensions are associated with low reliabilities. It appears that future research on the Barth Scale should include the creation of additional items for most scales to increase their levels of alpha internal consistency reliability.

### *Summary and Conclusions*

The judges tend to sort the Barth Scale items into general and easily identifiable content categories. In some cases, the response data, as compared with the Judgmental Data, reflected a more specific interpretation of factors. An example can be seen when one looks at judgmental Category 2 (Evaluation). The content experts sorted all the items pertaining to evaluation into that category. In contrast, the

response data suggested two aspects of evaluation: The Child (Factor III) and the Child's Work (Factor VI). It was suggested that the personal nature of the teacher-child relationship contributed to the teachers' interpretation of the items.

Thus, the results of the content validity study based on judgmental data gathered from content experts contributed to naming the dimensions derived from analyzing actual classroom teacher response data. Although facilitating the understanding of the constructs under study, the differences between the judgmental categories and response data dimensions do lead one to consider the possible disparity between theory and practice in open education. If disagreement between theorists and practitioners does exist, what are the resulting implications for open classroom implement? It appears that future research in open education might take this question into consideration.

### REFERENCES

- Barth, R. S. Open education: assumptions about learning and knowledge. Unpublished doctoral dissertation, Harvard University, 1970.
- Barth, R. S. So you want to change to an open classroom. *Phi Delta Kappan*, 1971, 53, 97-99.
- Bussis, A. M. and Chittenden, E. A. *Analysis of an Approach to Open Education*. Educational Testing Service, Princeton, N. J.: August, 1970.
- Gable, R. K. and Pruzek, R. M. *Methodology for instrument validation: an application to attitude measurement*. Paper presented at the American Educational Research Association annual meeting, Chicago, Ill.: April, 1972.
- Hofmann, R. J. The obliquimax transformation. Unpublished doctoral dissertation, State University of New York at Albany, 1970.
- Stanley, J. K-R (20) as the stepped-up intercorrelation of items. *National Council on Measurements Used in Education Yearbook*, 1957, 14, 78-92.
- Walberg, H. J. and Thomas, S. C. *Characteristics of open education: toward an operational definition*. Newton, Mass.: TDR Associates, 1971.
- Wiley, D. Latent partition analysis. *Psychometrika*, 1967, 32, 183-193.



## DO THESE CO-TWINS REALLY LIVE TOGETHER? AN ASSESSMENT OF THE VALIDITY OF THE HOME INDEX AS A MEASURE OF FAMILY SOCIO-ECONOMIC STATUS<sup>1</sup>

LOUISE CARTER-SALTZMAN<sup>2</sup> AND SANDRA SCARR-SALAPATEK  
University of Minnesota

WILLIAM B. BARKER  
University of Pennsylvania

In a study of 400 pairs of same-sex twins, ages 10-16, in the Philadelphia area, the Home Index was used as a measure of SES. Because there were numerous disagreements between co-twins, analyses were done of each item. The Home Index was then rescored using only the 10 items that reached a criterion level of 75% twin agreement. Two scoring methods were used on the 10-item Home Index: one treating "don't know" as equivalent to a blank (yes = 2, no = 1, don't know = 0) and the other being a variation on that scoring method by giving additional weight to "don't know" responses (yes = 2, no = 0, don't know = 1). By correlating the three scorings (the original scoring of 24 items and the two scorings of 10 items) of the Home Index with each other, with census tract data, and with five cognitive measures used in the study (Raven Standard Progressive Matrices, Peabody Picture Vocabulary Test, Columbia Mental Maturity Scale, Benton Revised Visual Retention Test, and a Paired-Associate Test), it was determined that the original Home Index was a more valid measure for white subjects than for black subjects. It could not, however, be recommended as a highly valid measure of SES in either group.

In a study of 400 pairs of same-sex twins in the Philadelphia area, an

<sup>1</sup> The research was supported by grants from the National Institute of Child Health and Human Development (HD06502) and the W. T. Grant Foundation. The authors gratefully acknowledge the assistance of Valerie Lindstrom in administering the measures and William Thompson in analyzing the data.

<sup>2</sup> Author's address: Institute of Child Development, University of Minnesota, Minneapolis, Minnesota 55455.



individual measure of socio-economic status, the Home Index (Gough, 1954, 1970) was administered to all of the children, ages 10 to 15 years. The Home Index is intended for use with older elementary, junior high, and high school students to assess family status and life style characteristics (Gough, 1949; 1971a; 1971b). It is reported to correlate highly with other measures of socio-economic status (SES) such as other status questionnaires and parental occupational levels (Gough, 1949, 1971a) and to predict college attendance (Gough, 1971b).

The present study assessed the validity of reports on family social status by comparing responses of two members from the same family, in this case co-twins. Co-twin agreement on factual items about their family is a particularly good criterion for evaluating the validity of an SES measure. If two children in the same family, especially siblings of the same sex and same age, do not agree on the information, then one must question the usefulness of information obtained in the scale.

### *Methods*

#### *Subjects*

Twin pairs were recruited by letter from a complete list of twins in the Philadelphia public schools and from parochial and suburban schools by newspaper articles and a television news broadcast. The final sample of twins tested included 399 pairs and two sets of triplets. Of the 175 black twin pairs, 157 attended Philadelphia public schools, 18 other Philadelphia schools, and none suburban schools. Of 224 white pairs, 89 attended Philadelphia public schools, 42 other Philadelphia schools, and 93 suburban schools.

The twins came from the greater metropolitan area to form a representative sample of families in the Philadelphia area. Median income and educational levels were computed for the census tracts from which the twins were drawn. The median values of family income in the samples of blacks and whites are very close to the median figures reported for the Philadelphia metropolitan area. For the twin sample, the whites' median income in 1970 was \$11,000, median education was 11.9 years; blacks' median income was \$7,910, median education was 10.2 years.

#### *Procedure*

About 10 pairs of twins were tested at the same time; they were seated in alternate chairs and rows in a large auditorium. Co-twins were separated into different small groups, each with an adult leader

who answered questions and guided them through the afternoon's assessments. There was no opportunity for co-twins to collaborate on the tests.

The Home Index and other psychological measures were presented on 35 mm. slides, synchronized with audio tapes of instructions and items read aloud. No reading skills were required. Answer sheets were specially formatted for each test with only the appropriate number of items and answer alternatives. Items and alternative responses were numbered and lettered in black, primary-school type for increased clarity. Instructions on the use of the answer sheets were given prior to the Home Index, and children's accuracy in use of the sheets was noted by the group leaders for the first two Home Index items. This method of administration did not decrease the reliabilities on any of the cognitive measures.

### *Measures*

*Home Index.* The 1954 version of the Home Index consists of 24 items, divided into four subscales derived by factor analysis: *social status* (8 items), *ownership* (10 items), *socio-civic involvement* (4 items), and *aesthetic* (2 items). Possible responses to the item questions were "yes," scored as 2 points, "no," scored as 1, and "don't know," scored as 0. Blanks were also given a 0 score. This scoring deviates from Gough's method which gave "yes" responses a score of 1, and "no" responses a score of 0. The authors included a "don't know" response to increase the validity of the data obtained.

Although in Gough's 1970 version of the Home Index the items "radio" and "television" were deleted, in the current form "radio" was retained, and "television" changed to "color television" in the hope that the discriminability of the item would be increased.

*Cognitive measures.* Five cognitive tests were administered to the sample as part of two sessions, each lasting about 1¼ hours: the (Raven) Standard Progressive Matrices, sets A, B, C, and D (Raven, 1958); the Peabody Picture Vocabulary Test (Dunn, 1959); the Columbia Mental Maturity Scale (Burgemeister, Blum, and Lorge, 1959); the Revised Visual Retention Test (Benton, 1963), and a paired-associate learning test (Stevenson, Hale, Klein, and Miller, 1968).

*Census tracts.* Median education and median income levels for every census tract in which a twin family lived were ascertained from the 1970 U.S. census. The distribution of income and educational levels was divided around the median to form higher and lower SES subgroups in each racial group. Mixed cases of above and below median values for income and education were assigned to the lower SES sub-

groups. Because the overlap in census tract medians was so small, it was impossible to use a common median for the two races. Thus, higher and lower SES subgroups were *not* comparable across race.

*Other measures.* Many other measures of dental and physical growth, personality and self-esteem, blood groups, taste preferences, and dermatoglyphics were obtained but are not reported in this paper.

### *Statistical Analyses*

To obtain measures of co-twin agreement, intraclass correlations (McNemar, 1962, p. 284) were computed. Interclass correlations of Home Index scores with cognitive measures and with census tract data were calculated for a sample comprised of one twin from each family.

Home Index records with six or more consecutive missing answers or ten or more total missing answers were deleted from the study. Fourteen of the 804 records were lost in this way. Only 28 of the remaining subjects had one or two missing answers, which were scored as blanks. A few additional subjects were lost to subsequent analyses because of incomplete data on other measures.

## *Results*

### *Twin Agreement*

Correlations of the total Home Index score of one twin with the co-twin revealed substantial disagreement on family information. The overall correlation for all twin pairs was only .72. Further analysis of subgroups indicated that disagreements were far more common for black pairs ( $r = .57$ ) than for white pairs ( $r = .77$ ), and that higher SES pairs in both races had somewhat less agreement than did lower SES pairs.

An analysis of agreement for each item was done to explore the sources of co-twin variance. Of the 24 items in the Home Index, only 10 were found to achieve 75% agreement in both racial groups, and only three items to achieve 90% agreement. The distribution of agreement (percentage of pairs where both twins gave the same response: "yes," "no," or "don't know") by subtest is found in Table 1. Male pairs and female pairs were considered separately.

Overall agreement was higher for white than for black pairs on all items. There were no consistent sex differences in agreement for the white pairs, but black male pairs were found to agree more often than were black female pairs for 18 of the 24 items, with two ties.

Using Gough's 22-item version (1970) in the analysis was con-

TABLE 1  
*Percentage of Co-Twin Agreement on Home Index Items*

	Black Pairs		White Pairs	
	Male (73)	Female (96)	Male (120)	Female (104)
<i>Subtest 1: Social Status</i>				
6. Did your mother go to high school?	75	67	87	86
7. Did your mother go to a college or university?	56	61	68	73
8. Did your father go to high school?	60	59	77	82
9. Did your father go to a college or university?	64	53	74	73
*10. Do you have a fireplace in your home?	86	79	91	93
*15. Does your family have any servants, such as a cook or maid?	81	79	90	89
16. Does your family leave town every year for a vacation?	60	59	69	84
23. Does your family have more than 500 books?	52	42	62	62
<i>Subtest 2: Ownership</i>				
*1. Is there an electric or gas refrigerator in your home?	93	84	96	97
*2. Is there a telephone in your home?	99	94	99	96
*3. Do you have a bathtub in your home?	95	93	98	98
4. Is your home heated with a central system, such as by a furnace in the basement?	59	59	73	78
*5. Does your family have a car?	89	85	98	98
*12. Does your family have a radio?	97	92	97	98
*13. Does your family have a phonograph (record player)?	85	81	87	86
20. Do you have your own room at home?	75	68	76	88
24. Does your family own its own home?	68	65	76	75
*14. Does your family have a color television at home? <sup>1</sup>	89	90	98	97
<i>Subtest 3: Socio-Civic Involvement</i>				
17. Does your mother belong to any clubs or organizations, such as . . .	67	66	73	76
18. Does your father belong to any civic, study, soc., or polit. clubs, . . .	64	70	68	70
21. Does your family subscribe to a daily newspaper?	62	60	81	90
22. Do you belong to any club where you have to pay dues?	68	74	72	75
<i>Subtest 4: Aesthetic Involvement</i>				
*11. Do you have a piano in your home?	89	88	98	96
19. Have you ever had private lessons in music, dancing, art, etc., outside of school?	66	68	75	73

\* Items that achieved 75% or more agreement.

<sup>1</sup> This item was changed from television to color television because the original item had no correlation with social status.

sidered, but since "radio" and "color television" were the third and fourth most reliable items, respectively, they were retained for all analyses.

Since it is possible that twin agreement might increase as a function





intercorrelations of the three Home Indices and their correlations with the criterion measures.

The three scorings for the Home Index (the original scoring of 24 items and the two different scorings of ten items) gave results that were positively intercorrelated, as expected. The  $10_1$  and  $10_2$  versions were highly related, especially for the white pairs. The Home Index measures did not correlate so highly with the cognitive tests as did the census tract median data in the black group. In the white group, the three Home Indices did correlate slightly more highly with the cognitive tests than did the census tract data. Although the three Home Indices were positively correlated with census tract data in both racial groups, the coefficients were not very high.

Co-twin agreement by race and SES (census tracts) for the Home Indices is given in Table 3. Home Index  $10_2$  is clearly superior in degree of co-twin agreement to the other two versions of the scale in both black and white groups and for the lower and higher SES pairs within both races. Agreement between co-twins was lowest on the  $10_1$  version and highest on the  $10_2$  version. This result indicated that "don't know" was an intermediate response, given when one twin had information (yes or no) and the other did not.

One possible explanation of the low correlations between the rescored Home Indices and the criterion measures could be the reduced variance of scales with ranges of 0-20 instead of the original 0-48 range of the 24-item Home Index. However, it was found that Home Index  $10_2$  did have a higher variance than did  $10_1$  but correlated no more highly with the criterion measures. It is, therefore, doubtful that reduced variance could account for the correlational results.

It is also possible that the 10 valid items did not discriminate to a high degree between groups because of the marked unevenness of frequency distributions of responses over item alternatives. For example, nearly everyone has a bathtub, refrigerator, and radio, but few have servants or a piano. There are several items, however, for which the frequencies of possession and nonpossession are relatively equal (color TV, fireplace, telephone, car, and phonograph).

### *Race and Age Differences*

Overall, the white co-twins agreed more often than black co-twins on family information. With a 90% agreement criterion, eight items qualified as valid for the white group, whereas only three items were valid for blacks (see Table 1). A 75% criterion selected 15 of the 24 items for both sexes of white twins and only 10 items for the blacks. White twins did agree more often with each other than did black twins

TABLE 3

*Twin Correlations for Three Versions of the Home Index by Race and Race and SES (Census Tracts)*

	Lower (94)	Black Higher (78)	All (172)	Lower (124)	White Higher (97)	All (221)	Total
Home Index 24	.59	.44	.57	.79	.65	.77	.72
Home Index 10 <sub>1</sub>	.15	.35	.28	.60	.79	.71	.49
Home Index 10 <sub>2</sub>	.48	.70	.60	.79	.88	.85	.74

on parental education, home ownership, newspaper subscription, and having their own bedrooms.

Older twin pairs tended to agree more often than younger pairs, especially in the white group. In the black group, however, disagreements were far more frequent and less related to age than in the white group. In the white group the validity of information received from the Home Index clearly increased with age—a finding which suggests that a restriction of use of the Home Index to white high school students would increase its validity. In the black group, for whom increased age did not improve the validity of information given, the Home Index was probably not a useful instrument.

#### *Toward SES Measurement*

The criterion measures of median census tract education and income levels and the five cognitive scores correlated more highly with the rescored Home Indices in the white than in the black group. Gough (1971a) reported a correlation of .21 between the original Home Index and intellectual ability in a high school population. The correlations between the cognitive measures and (a) the original set of scores and (b) each of the two sets of scores for the 10-item Home Index were of similar magnitude in the white twin sample.

The size of mean differences between the white and black groups on the Home Index measures was small, as compared to that of the differences in the census tract data. On the Home Index, 18.5% of the black children equalled or exceeded the median for the white children. The extent of census tract overlap was considerably less: for education about 8% of the blacks were equal to or exceeded the white median, and for income about 2% of the blacks were at or above the median of the whites. The Home Index appeared to minimize estimates of SES differences between the two racial groups and within each racial group.

The lack of co-twin agreement on many of the Home Index items

raises serious questions about its usefulness as a measure of individual SES characteristics for a 10- to 16-year-old group. That the Home Index failed to correlate with intellectual measures more highly than with census tract data and that it failed to discriminate clearly between disadvantaged and advantaged populations would suggest that its validity is in doubt. Since ownership items proved to be the most reliable, it is recommended that additional items, preferably items that would afford alternative responses at intervals within the middle range of the SES distribution, be added to a revised scale, and that data be collected from two informants in each family to substantiate the validity of a new scale with 10- to 16-year-old subjects. Parental reports of SES characteristics would be a particularly valuable comparison for the reports of their children.

### REFERENCES

- Benton, A. L. The Revised Visual Retention Test, Form C. Dubuque, Iowa: William C. Brown, 1963.
- Burgemeister, B. B., Blum, L. H., and Lorge, I. Columbia Mental Maturity Scale. New York: Harcourt, Brace, and World, 1959.
- Dunn, L. M. Peabody Picture Vocabulary Test. Circle Pines, Minnesota: American Guidance Service, 1959.
- Gough, H. G. A short social status inventory. *Journal of Educational Psychology*, 1949, 13, 534, 537.
- Gough, H. G. The home index. Berkeley: University of California, 1954.
- Gough, H. G. The home index. Berkeley: University of California, 1970.
- Gough, H. G. A cluster analysis of home index status items. *Psychological Reports*, 1971, 28, 923-929 (1971a).
- Gough, H. G. Socio-economic status as related to high school graduation & college attendance. *Psychology in the Schools*, 1971, 7, 226-239 (1971b).
- McNemar, Q. *Psychological statistics* (3rd ed.). New York: Wiley, 1962.
- Raven, J. C. Standard Progressive Matrices: Sets A, B, C, D, & E. London: Lewis, 1958.
- Stevenson, H. W., Hale, G. A., Klein, R. E., and Miller, L. K. Interrelations and correlates in children's learning and problem solving. *Monographs of the Society for Research in Child Development*, 1968, 33, Whole No. 123.



## STABILITY OF STUDENT EVALUATIONS OF INSTRUCTORS AND THEIR COURSES WITH IMPLICATIONS FOR VALIDITY

HENRY J. OLES

Southwest Texas State University

A course-instructor evaluation form was administered to 775 undergraduates in 15 large and small section introductory courses after the second class meeting and again near the end of the semester. The median pretest posttest correlation was  $+ .60$ . Students were generally more negative toward their course and instructor at the end of the semester than they were at the beginning.

As a separate portion of this project, two instructors deliberately attempted to alter their students' evaluation in one of two large sections of their introductory psychology course. In both cases, there was a significant overall mean difference between the experimental and control groups on the initial evaluation but there was no difference on the end-of-semester evaluation.

The results of this study indicated that although students quickly form reasonably lasting judgments of their instructors and courses they are also able to alter their judgments as warranted by changing situations. These findings appear to provide support for the validity of student evaluations.

THE use of student evaluations of faculty and courses is now common on most college campuses with the evaluative information being used both by students and administrators for decision making. Although many arguments have been made against using student evaluations as a primary criterion for professional advancement (Dressel, 1973), most of the research findings have indicated that student evaluations are reliable and reasonably valid indicants of teacher performance. Regardless of the opinion of academia, student evaluations are being used at most institutions of higher education for a wide variety of purposes. Therefore, it is of prime importance to continue to



conduct research on student evaluations to determine and improve their reliability, validity, and utility.

Several recent studies have attempted to determine the relationship between ratings made while the course was in progress with those made at the end of the course (Dick, 1967; Costin, 1968.) Bausell and Magoon (1972) reported a median correlation of .67 between ratings made at the end of the first class period with those made at the end of the semester. This finding was of particular concern to this researcher, since it could indicate that students enter a course with a definite predisposed and unalterable set of feelings about the course and instructor or that they quickly form a rigid and lasting set of attitudes after only minimal exposure.

### *Purpose*

This study was designed to replicate and expand upon the work of Bausell and Magoon concerning stability of ratings of selected teacher and course characteristics from the beginning to the end of a semester. Within the college and university setting such information was thought to be of importance in judging the potential validity of an evaluation of a course and teaching performance at two widely separated time points. Bausell and Magoon used upper level undergraduates and graduate classes with a median size of 15. In addition they emphasized a standard student evaluation form for collecting both beginning and end-of-course evaluations. In a pilot study this researcher has found that undergraduate students vehemently objected to evaluating a teacher or course after the first or second class day when they were required to complete a form that obviously had been designed for use at the end of the semester. Therefore, the form used in this study was designed to overcome student objections through careful wording of the directions to the respondent on the pretest as well as through emphasis of the fact that the form was specifically designed to measure their *first impression* of the instructor and his course. In addition, each of the questions was worded to make it appropriate for a first impression evaluation. The posttest was essentially the same as the pretest with only minor changes in tense (i.e., whereas the pretest stated "This teacher seems to be . . ."; the posttest stated "This teacher was . . .").

Virtually all research on student evaluations has been conducted after the fact. This researcher and a colleague each taught two essentially identical sections of introductory psychology with approximately 125 students in each section. A deliberate attempt was made to create a negative first impression in the experimental section by begin-

ning the course with an unusually dry lecture on the historical roots of the science of psychology and on the methods of science to determine whether this treatment would alter the student ratings as compared with those of a class receiving a high interest-arousing presentation. If a variation in instructor performance was reflected in student ratings in the direction intuitively expected, the result would add to the construct validity of student ratings in general, since many skeptics have insisted that student ratings are not directly related to any meaningful teacher behavior.

### *Methodology*

The subjects for this study included 1302 undergraduate students enrolled in 15 lower division courses taught by 13 different instructors with class sizes ranging from 17 to 154.

The scale consisted of 22 evaluative items (21 on the pretest) covering various dimensions of instructor performance and of the course. Each item consisted of a statement describing the instructor or the course followed by four or five evaluative phrases ranging from very positive to very negative. The overall rating was composed of a simple summation of the individual ratings. Lower numerical ratings indicated the more positive attitudes.

During the second class meeting, all students in 15 undergraduate courses were asked to complete the first impression instructor-course rating form. The instructor was asked to leave the room while the forms were distributed. Each group was told that the purpose of the first impression rating form was to help improve the design of student evaluation forms. They were reminded several times that their instructor would not see the ratings until after the semester was complete and grades were submitted. Essentially identical instructions were given with the posttest which was administered by the same person during the last week of the semester. The students were not aware of the posttest when they took the pretest. An identification system was used to enable the matching of pre and post evaluations that still permitted respondents to remain anonymous.

Two of the 13 instructors involved in the project each taught two essentially identical sections of introductory psychology. Their normal approach to beginning the introductory course was quite different. One instructor (instructor A) used several interest-arousing lectures while the other (instructor B) plunged in the first day with an admittedly dry, at least in terms of student interest, lecture on the methods of science and historical perspectives in psychology. Each instructor agreed to attempt to alter his behavior in one class to match

that of his colleague. This procedure resulted in two classes that received a high interest introductory lecture and two classes that received a rather low interest lecture. The instructors were then told to continue the semester after the second day with their standard style of teaching. Ideally, this portion of the study should have been extended to a significantly larger number of instructors and courses. However, this researcher was concerned about the moral and ethical obligations of every teacher to do his best in teaching his courses. Therefore, the decision was made to use deliberate modification of normal teaching practice in only two highly controlled situations even though this decision would result in some questioning of the validity and generalizability of the findings.

Two threats to the internal validity of the procedures employed in this investigation were (a) the possible effects of having taken the pretest on posttest performance and (b) the students' familiarity with the instructor before the first class meeting. Bausell and Magoon specifically examined their data for pretest sensitization and found none. No similar test was performed in this study; however, observation of student reactions to the posttest indicated that they had virtually forgotten having taken the pretest three months previously. None of the students had had any previous classroom exposure to the instructor, since all the courses were introductory.

### *Results and Interpretation*

#### *Pretest-Posttest Comparisons*

Of the 1320 students who took the pretest, 775 were matched with their posttest ratings. Approximately 40% of the subjects was lost because of absences, withdrawals, incomplete forms, and inability to match the two forms.

Table 1 presents the percentage of subjects selecting each response option for 21 items on the pretest and posttest and for one item found only on the posttest. The most interesting finding is the large proportion of students who chose the most favorable response options, 0 and 1. Response option 2 was rarely selected for most items while option 3 responses were essentially nonexistent, especially on the pretest. Students were evidently inclined to give positive ratings even to relatively poor teachers.

The mean ratings for each of the evaluative items was calculated for each class on the pretest and posttest. Pretest posttest mean ratings were significantly different at the .05 level for nine items. As compared with their average pretest standing, the average posttest standings of

TABLE 1  
*Percentage of Students Selecting Each Response Option*

Item	Response Option Pretest					Response Option Posttest				
	0	1	2	3	4	0	1	2	3	4
1. Interest in Course	40	48	9	2	0	28	44	18	7	2
2. Course Difficulty	4	39	47	9	0	8	43	40	9	1
3. My Grade	22	62	15	0	0	9	40	42	9	1
4. Textbook	17	47	32	5	0	17	36	32	12	4
5. Course Organization	39	57	3	0		40	53	6	1	
6. Teachers Knowledge	79	21	0	0		78	21	1	0	
7. Teachers Attitude Toward Course	67	30	2	1		64	32	3	1	
8. Teachers Explanations	58	37	5	0		50	39	9	1	
9. Intellectual Stimulation	24	66	10	0		19	60	20	1	
10. Speaking Ability	68	30	2	0		63	33	3	1	
11. Teachers Attitude Toward Students	53	30	16	1		56	32	11	1	
12. Grading Fairness	18	79	3	0		53	42	4	1	
13. Tolerance to Disagreement	58	39	2	0		54	41	3	2	
14. Teachers Personality	59	36	3	2		57	38	3	2	
15. Overall Rating	20	47	32	2		24	47	25	4	
16. Desire to Attend Class	58	40	1	1		37	54	7	2	
17. Value of Attendance	96	4	1			81	15	4		
18. Utilization of Time	80	19	1	0		70	25	4	1	
19. Amount Learned	64	34	2			47	45	8		
20. Satisfaction With Course	70	26	4			77	17	6		
21. Sticks to Subject	66	32	2			61	35	4		
22. Recommend to Friends (posttest only)						58	29	10	3	

the students indicated significantly less interest in their course, expectation of a lower grade, greater objectionableness to the textbook, finding explanations by the teacher more inadequate, less desire to attend class, seeing less value in attending class, and wasting of more class time by the instructor at the end of the semester than at the beginning. However, students did see examinations and grading as being more nearly fair at the end of the course than at the beginning even though many had expected to receive considerably lower grades than they had initially anticipated.

The median pre-posttest correlation for all 21 items was .60 and ranged from  $-.11$  for the amount of information learned to  $+.86$  for course difficulty and attractiveness of the teacher's personality. Generally, the obtained correlations could be rationalized. Those aspects of the course that could potentially be reliably and validly assessed at the beginning were highly correlated with posttest ratings, whereas those aspects that could conceivably be accurately rated only after several weeks of exposure showed low correlations.

Table 3 presents the results of a deliberate attempt by two instruc-

TABLE 2  
Means<sup>a</sup>, Standard Deviations, *t* Tests and Correlation between Pre- and Posttest Mean Ratings

Item	Pretest <i>M</i>	SD	Posttest <i>M</i>	SD	Difference	<i>t</i>	<i>r</i> ( <i>N</i> = 15)
1. Interest in Course	.84	.40	1.23	.59	+.39	3.32**	.64
2. Course Difficulty	1.52	.32	1.43	.29	-.09	.79	.86
3. My Grade	.95	.19	1.58	.28	+.63	11.29**	.63
4. Textbook	1.25	.25	1.70	.57	+.45	4.14**	.75
5. Course Organization	.68	.15	.73	.29	+.05	.84	.60
6. Teachers Knowledge	.22	.11	.24	.15	+.02	.78	.63
7. Teachers Attitude Toward Course	.42	.33	.44	.33	+.02	.30	.72
8. Teachers Explanations	.51	.21	.69	.35	+.18	2.15*	.42
9. Intellectual Stimulation	.92	.22	1.03	.22	+.11	1.50	.20
10. Speaking Ability	.40	.26	.49	.34	+.09	.83	.71
11. Teachers Attitude Toward Students	.71	.37	.62	.29	-.09	1.14	.57
12. Grading Fairness	.82	.12	.58	.25	-.24	-4.47*	.63
13. Tolerance to Disagreement	.54	.37	.56	.25	+.02	.19	.18
14. Teachers Personality	.63	.44	.64	.41	+.01	.24	.86
15. Overall Rating	1.21	.46	1.16	.45	-.05	.44	.60
16. Desire to Attend Class	.43	.15	.71	.25	+.28	5.46**	.60
17. Value of Attendance	.20	.42	.33	.42	+.13	2.04*	.83
18. Utilization of Time	.19	.12	.33	.22	+.14	2.79*	.53
19. Amount Learned	.49	.25	.65	.32	+.16	1.64	.11
20. Satisfaction With Course	.43	.30	.37	.33	-.06	.74	.53
21. Sticks to Subject	.34	.23	.41	.19	+.07	1.19	.33
Total Overall Rating—All 21 Scales Combined	.62	.37	.72	.40	+.10	2.70*	.90

\* Significant at .05.

\*\* Significant at .01.

<sup>a</sup> Lower Mean = More Positive Rating.



TABLE 3

*Comparison of Pre- and Posttest Ratings When Instructors Deliberately Altered Their Normal Teaching Style*

INSTRUCTOR A					
Pretest (interest arousing)	Pretest (no interest)		Posttest (interest arousing)	Posttest (no interest)	
$M = .48$	$M = .63$	$t = 2.83^*$	$M = .58$	$M = .59$	$t = .33$
$SD = .36$	$SD = .33$	$r = .80$	$SD = .36$	$SD = .39$	$r = .98$
INSTRUCTOR B					
Pretest (interest arousing)	Pretest (no interest)		Posttest (interest arousing)	Posttest (no interest)	
$M = .60$	$M = .95$	$t = 4.97^{**}$	$M = 1.16$	$M = 1.20$	$t = .72$
$SD = .40$	$SD = .42$	$r = .73$	$SD = .52$	$SD = .48$	$r = .92$

\* Significant at .02 level.

\*\* Significant at .01 level.

tors to alter their initial expected student ratings in two of the four essentially equivalent sections of introductory psychology. Both instructors used a highly interesting introductory lecture in one section and a rather dry monotone lecture in the other section. The difference between the mean pretest ratings for Instructor A were significant at the .02 level and for Instructor B, beyond the .10 level. That there were no statistically significant differences on the posttest ratings for either instructor indicated that the students were indeed able to alter their first impression ratings to fit the instructors typical performance shown throughout the semester. Although the differences in mean ratings between the interest and noninterest arousing introductory lectures were highly significant, the generalizability of this finding is low because of the small  $N$  (2 instructors, 4 sections). However, this researcher believes that these findings are of critical importance in demonstrating the validity of student evaluations. This portion of the study demands replication on a larger scale, if adequate control can be maintained to protect those students who may be inadvertently negatively affected by unknowingly being part of the experimental group.

### *Summary and Conclusions*

This investigation examined three aspects of student rating of college instructors; (a) the distribution of ratings given after the first or second class meeting and again during the last week of the semester; (b) the correlation between mean pretest and posttest ratings for each item using fifteen classes; and (c) the effects of short-term deliberate manipulation of teaching style on ratings.

The data in Tables 1 and 2 reveal that students in this study had a definite tendency to rate instructors positively on both the pretest and posttest, although ratings on the posttest were generally more negative than those on the pretest (and variable) as shown by the higher rating values assigned. Only one item, expected course difficulty, had ratings above the expected mean (1.5) on the pretest. The overall mean for the combined 21 items on the pretest and posttest were .62 and .72, respectively. Those institutions that passively permit some use of student evaluations without offering individual faculty members a statistical analysis of their ratings with respect to those of other members of the department, school, or institution, may in actuality be promoting a false sense of satisfaction and security among faculty, since individual faculty members may not be aware of the students tendency to report above average ratings. Obviously the reliability and differential validity of student evaluations would be improved if techniques were used to encourage students to rate realistically the relative effectiveness of their teachers on a true four point scale.

The median correlation between beginning and end of semester ratings was shown to be +.60 with individual item correlations ranging from -.11 for amount of material learned to +.86 for assessment of the course difficulty and the teacher's perceived personality. The results of this portion of the study demonstrated that students were able to form relatively lasting appraisals of their course and instructor after minimal exposure. The stability of the ratings listed in Table 3 met logical expectations. Although all characteristics of a course can be misjudged, those particular characteristics that would be anticipated to require maximum exposure in order to make a realistic judgment, indeed, showed the lowest pretest-posttest correlations (Learned -.11; Tolerance to disagreement, .18; Intellectual stimulation, .20).

The final portion of this study was designed to determine whether or not students in experimental and control groups would give significantly different ratings to teachers who had altered their teaching style in two introductory psychology courses. Information in Table 3 shows that indeed students rated the two styles of teaching differently. The life orientated, interest arousing approach did receive significantly average higher mean ratings,  $t = 2.83$  and  $4.97$  respectively, than did the noninterest arousing, basic science/historical approach. Nearly all individual ratings were more negative for the rigid non-interest approach in both experimental groups. There were no significant differences in the mean ratings at the end of the semester between the experimental and control groups for each instructor.

That the correlations between the mean item ratings in the experi-

mental and control groups on the posttest for instructors A and B were .98 and .92, respectively, would suggest a relatively high degree of reliability for the rating instrument across subjects.

The generalizability of this portion of the study, however, is questionable because of the participation of only two instructors which was a result of this researcher's concern for maintaining as high a degree of control as possible for the ethical responsibility of an instructor to do his best in a course. However, because of the highly significant results reported and because of their importance to establishing experimentally the validity of student evaluations, this portion of the project should serve as a pilot study to be replicated subsequently on a larger scale.

### REFERENCES

- Bausell, R. B. and Magoon, J. The persistence of first impressions in course and instructor evaluation. Unpublished paper, American Educational Research Association, 1972.
- Costin, F. A graduate course in the teaching of psychology: Description and evaluation. *Journal of Teacher Education*, 1968, 19, 425-432.
- Dick, W. *Course attitude questionnaire: Its development, uses, and research results*. University Division of Instructional Services, The Pennsylvania State University, Report No. 106, revised by D. Stickell, September, 1967, (mimeographed).
- Dressel, P. Student evaluation of faculty. Why? What? How? In A. L. Sockloff (Ed.) *Proceedings, The First Invitational Conference On Faculty Effectiveness As Evaluated By Students*. Philadelphia: Measurement and Evaluation Center, Temple University, 1973.



## PROTESTANT ETHIC ATTITUDES AMONG COLLEGE STUDENTS

L. K. WATERS, NICK BATLIS, AND CARRIE WHERRY WATERS  
Ohio University

The six scales of the Survey of Work Values (Wollack, Goodale, Wijting, and Smith, 1971), the Blood (1969) pro-Protestant Ethic scale, and the Protestant Ethic scale of Mirels and Garrett (1971) were intercorrelated, and each scale was correlated with Rotter's I/E scale, SAT total score, and cumulative grade point average for 170 college students. A factor analysis of the Protestant Ethic scales yielded two factors which were interpreted, on the basis of the loadings of the Survey of Work Values scales, as representing intrinsic (work-related) and extrinsic (reward-related) aspect of Protestant Ethic. The Blood and the Mirels and Garrett scales loaded substantially on both factors. Generally, the Protestant Ethic scales were negatively related to external orientation on the I/E scale, and were unrelated to SAT scores and academic performance.

RELATIVE to the Protestant Ethic, Wollack, Goodale, Wijting, and Smith (1971) have recently stated:

The principal aspects of the Protestant Ethic as described by Weber (1958) are individualism, asceticism, and industriousness. The emphasis placed on a man's industriousness probably represents the most critical aspect of Protestant Ethic. The Ethic has been assessed typically by indirect methods that have been presumed to index this concept . . .

Correlations between attitudes and supposedly logically related behaviors have usually been found to be low. Frequently, a variety of considerations intervene to inhibit the behavioral manifestations of an attitude. Economic and social factors may greatly limit the alternative behaviors available to an individual regardless of his attitudes. It would, therefore, be naive to expect indirect measures to index accurately an individual's work values. Attitude scales



should provide more direct measures of these concepts [p. 331].

Within the past few years three scales have been developed to measure Protestant Ethic attitudes, and validity data have been reported for each of the scales (Blood, 1969; Mirels and Garrett, 1971; Wollack, Goodale, Wijting, and Smith, 1971).

The purpose of the present study was two fold: (a) to determine, for a college sample, the interrelationships among the three Protestant Ethic scales, and (b) to examine the relationships of each of the three scales to selected personality, ability, and academic performance measures.

### *Method*

All scales were administered in booklet form to 102 males and 63 females enrolled in introductory level psychology classes. Items from the three Protestant Ethic scales were combined into one section of the booklet, and each item was responded to on a 7-point agree/disagree scale.

One Protestant Ethic scale (PE-B) consisted of the four pro-Protestant Ethic items used by Blood (1969) in a study of work values and job satisfaction among military personnel. The second Protestant Ethic scale (PE-MG) was that developed by Mirels and Garrett (1971) with samples of college students. This scale consists of 19 items. Each of these scales yields a single overall score. The third Protestant Ethic scale was the Survey of Work Values (SWV) developed by Wollack, et al. (1971). The SWV has 54 items with nine items covering each of the six areas: Activity Preference (AP), Job Involvement (JI), Pride in Work (PW), Social Status of Job (SS), Upward Striving (US), and Attitude toward Earnings (AE). The first three subscales of SWV represent three dimensions of Protestant Ethic that cover intrinsic aspects of work; whereas the latter three subscales are intended to reflect extrinsic aspects of the Protestant Ethic.

The personality measure included in the booklet was Rotter's (1966) internal/external control of reinforcement scale (I/E). The two other indices were taken from student records: the total score of the Scholastic Aptitude Test (SAT) of the College Entrance Examination Board, as a measure of ability, and cumulative grade point average (GPA) as a measure of academic performance.

### *Results and Discussion*

The correlations of the Protestant Ethic scales with each other and with the personality, ability, and academic performance measures are

given in Table 1. Coefficient Alpha reliability estimates are presented in the diagonals for the Protestant Ethic scales.

The extrinsic subscales of the SWV had internal consistency reliability estimates similar to and the intrinsic subscales had reliability estimates somewhat higher than those reported by Wollack, et al. (1971) for industrial and government workers. The PE-MG scale's reliability was almost identical to that reported by Mirels and Garrett (1971) for the college students. Blood (1969) did not report a reliability estimate for his pro-Protestant Ethic scale, but the obtained .71 coefficient seems quite sufficient for research purposes (and quite high considering the scale has only four items).

A principal components factor analysis of the correlations among the Protestant Ethic scales (using 1.00 as the eigenvalue cutoff and a Varimax rotation) yielded two factors which were consistent with the two-factor intrinsic (work-related) and extrinsic (reward-related) dichotomy. On the first factor, loadings of .85, .77, and .89 were obtained for the SWV intrinsic subscales AP, JI, and PW, respectively. The SWV extrinsic subscales SS (.77), US (.70), and AE (.80) defined the second factor. Both the Blood and the Mirels and Garrett scales, loaded in the middle to high .50's on both factors.

The intrinsic subscales of the SWV, the PE-B, and the PE-MG scales all correlated significantly ( $p < .01$ ) and negatively with the I/E scale—an outcome indicating that persons scoring high on these PE scales tended to perceive their own efforts and abilities, rather than luck or fate, as determining the course of events in their lives. This result is consistent with the findings of Mirels and Garrett (1971) for college students.

TABLE I  
*Correlations of Protestant Ethic Scales with Each Other and Measures of Personality, Ability, and Academic Performance<sup>a</sup>*

Measures <sup>b</sup>	AP	JI	PW	SS	US	AE	PE-B	PE-MG	I/E	SAT	GPA
AP	(.77)	.52*	.71*	.15	.24*	-.11	.49*	.49*	-.28*	.03	.20
JI		(.69)	.64*	.20	.34*	-.11	.35*	.37*	-.33*	.00	-.06
PW			(.76)	.14	.36*	-.15	.45*	.47*	-.29*	.02	.11
SS				(.64)	.50*	.45*	.36*	.35*	-.16	-.15	-.03
US					(.63)	.34*	.48*	.47*	-.18	-.21	-.07
AE						(.62)	.27*	.26*	.03	-.28*	-.16
PE-B							(.71)	.70*	-.35*	.03	-.09
PE-MG								(.80)	-.34*	.08	-.12
$\bar{X}$	45.64	48.42	52.99	30.48	39.39	30.87	17.19	81.99	12.09	946.89	2.86
SD	7.27	6.61	6.76	7.36	6.68	6.82	4.77	13.89	4.44	162.62	.54

<sup>a</sup> Decimal points omitted from the correlation coefficients.

<sup>b</sup> The names and descriptions of the measures are presented in the text.

\*  $p < .01$ .

With the exception of the SWV extrinsic subscales, the PE scales were almost completely unrelated to ability as measured by SAT total score. Although all three SWV extrinsic subscales were negatively related to SAT, only AE correlated significantly ( $p < .01$ ). None of the PE scales was significantly related to academic performance indexed in terms of cumulative GPA (with or without SAT total partialled out). Mirels and Garrett (1971) found, in terms of interest patterns on the Strong Vocational Interest Blank, that PE attitudes were positively related to occupations requiring a concrete, pragmatic approach to work and negatively to SVIB scales for occupations which to a greater degree require emotional sensitivity, theoretical interests, and humanistic values. With the wide range of contemplated majors in a relatively unselected group of freshman and sophomore students, it seems reasonable that Protestant Ethic scales did not correlate with GPA. Perhaps PE scales would show differential correlations with performance for specific majors.

## REFERENCES

- Blood, M. R. Work values and job satisfaction. *Journal of Applied Psychology*, 1969, 53, 456-459.
- Mirels, H. L., and Garrett, J. B. The Protestant ethic as a personality variable. *Journal of Consulting and Clinical Psychology*, 1971, 36, 40-44.
- Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, 1966, 80, (1 Whole No. 609).
- Weber, M. *The Protestant ethic and the spirit of capitalism*. New York: Scribner, 1958.
- Wollack, S., Goodale, J. G., Wijting, J. P., and Smith, P. C. Development of the survey of work values. *Journal of Applied Psychology*, 1971, 55, 331-338.

## PREDICTIVE VALIDITY OF THE AMERICAN UNIVERSITY OF BEIRUT TRIAL APTITUDE BATTERY

F. K. ABU-SAYF AND GEORGE I. ZA'ROUR

American University of Beirut

On the basis of student scores on a battery of tests given at the time of their admission to the American University of Beirut and their subsequent college grades, the validity of the three new tests for predicting college grades was obtained. The results in general suggested a low to moderate predictive validity for each of the tests. Suggestions for test revision were based on item analysis and on classification of the items in terms of both the factors of science aptitude and the subject matter each item represents.

In the fall of 1971, two essentially unspecced 50-item, 5-choice forms of each of three new tests were administered to newly enrolled students at the American University of Beirut (AUB). These tests were: the English Proficiency Test (EP), the Quantitative-Aptitude Test (AQ), and the Science-Aptitude Test (AS). The purpose of this investigation was to find the predictive validity of each of the tests and of weighted combinations of them with grade-point average (GPA) and performance in selected courses as criteria. Special emphasis was given to AS in view of its particular relevance to science curricula that are emphasized at AUB.

### *Methodology*

For the most part, zero-order and multiple correlation coefficients were calculated. The experimental sample consisted of 275 freshmen and sophomores (except as otherwise noted).

### *Results*

#### *The Test Battery in General*

The intercorrelations among scores on the three tests showed that AS overlapped with AQ to a relatively large extent and with EP to a

smaller extent, whereas the overlap between EP and AQ was only moderately high. These values, like all the correlation coefficients computed in this project, were not corrected for restriction of range.

Product-moment correlation coefficients between scores on these tests and selected performance criteria suggested that the prediction of first-semester GPA was generally low to moderate, although the coefficients of correlation were generally significantly different from zero. AS predicted grades in Biology 201 more nearly accurately than grades in other science courses; AQ was a valid predictor of grades in mathematics courses; and EP predicted grades in English 201 at a higher level of accuracy than in other English courses.

Multiple correlation coefficients between possible combinations of these tests and first-semester freshman and sophomore GPA are presented in Table 1. These values were not appreciably higher than were the zero-order correlations of single tests with the same criterion. When first-year GPA was used as the criterion (instead of first-semester GPA), the data shown in Table 2 were obtained. Since cross-validation data were not available, shrinkage in these values would be expected to occur, especially where the numbers of cases were small. The underlined values were recommended as most efficient for practical considerations. It is also interesting to note the appreciable drop in the coefficients whenever EP was absent from a certain combination.

The point-biserial correlation coefficients obtained from item-analysis of each test using first-semester GPA as criterion were low. Consequently, the minimum value for an item to be left unchanged was taken to be 0.15.

TABLE 1  
*Multiple Correlation Coefficients between Either First-Semester Freshman or First-Semester Sophomore GPA as Criterion Variables and Combinations of Scores on EP(1), AQ(2), and AS(3) as Predictor Variables*

Predictor Variables	Subjects	N	R <sup>a</sup>
1,2	Freshmen	73	.35
	Sophomores	202	.37
1,3	Freshmen	73	.33
	Sophomores	202	.33
2,3	Freshmen	73	.28
	Sophomores	202	.30
1,2,3	Freshmen	73	.35
	Sophomores	202	.38

<sup>a</sup> All coefficients significant beyond the .01 level.



TABLE 2

*Correlation Coefficients between First-Year GPA(1) and EP(2), AQ(3), and AS(4) and Combinations of These Predictor Variables*

Correlation Coefficient	Arts (N = 75)	Bus. Adm. (N = 24)	Science (N = 174)	Total Sample (N = 271)
$r_{12}$	.57**	.50*	.56**	.55**
$r_{13}$	.05	.38	.56**	.33**
$r_{14}$	.08	.04	.46**	.33**
$R_{1.23}$	.58**	.53	.66**	.58**
$R_{1.24}$	.58**	.51	.60**	.57**
$R_{1.34}$	.08	.43	.62**	.37**
$R_{1.234}$	.59**	.57	.67**	.58**

\*  $p < .05$ .

\*\*  $p < .01$ .

Note—The underlined values denote recommended test combinations.

### *The Science-Aptitude Test in Particular*

Based on the Kuder-Richardson formula (21), the coefficients of internal consistency of the two forms of AS, coded 885 and 886, were found to be .81 and .80, respectively. Form 886 proved to be a slightly more valid predictor of GPA than was Form 885. Critical ratios between means suggested that science students did perform significantly better than did arts students ( $p < .01$ ). Form 886 was more difficult than Form 885 ( $p < .01$ ). Ten items out of 50 were found by inspection to be equivalent in both forms, whereas four pairs were equivalent within Form 886.

*Content validity.* The items in AS were classified in eight categories judged by previous investigators to define science aptitude: (1) tendency to suspend judgment when evidence is insufficient; (2) ability and accuracy in designing and defining; (3) creativity; (4) previous scientific experience; (5) an experimental bent; (6) specialized curiosity; (7) accuracy in reasoning and interpreting data as in the ability to evaluate and detect inconsistencies; and (8) mechanical ability. The items were then further classified under chemistry, physics, biology, physical chemistry and chemical physics (nuclear and atomic chemistry and physics), or general science (astronomy and space and earth sciences). As a result of this work, the items in each form of AS were arranged in a presumed order of decreasing validity. This ordering was used in revising the tests.

*Comments regarding the potential validity of AS.* The maximum correlation that could be attained between scores on any predictor test ( $t$ ) and scores on any criterion ( $c$ ) would be  $\sqrt{r_{tt}r_{cc}}$ , in which  $r_{tt}$  is estimated reliability of the test and  $r_{cc}$  is the estimated reliability of the criterion variable. If the test yields perfectly reliable scores and if the

criterion variable (in this case GPA) is expressed as having an average estimated reliability coefficient of about .81, the maximum validity coefficient that could be obtained is .90. This value is the upper limit for the predictive validity of Test AS which would be attained if the test were perfectly reliable and measured all the true variance of GPA. In practice, the validity of the test could most readily be increased by selecting for use, from item analysis procedures, those items which were found to be correlated *highest* with the external criterion variable and *lowest* with total test scores (Gulliksen, 1950, p. 380 ff.).

#### REFERENCE

Gulliksen, H. *Theory of mental tests*. New York: Wiley, 1950.

## VALIDITY OF AWARDING COLLEGE CREDIT BY EXAMINATION IN MATHEMATICS AND ENGLISH<sup>1</sup>

CAROL KEHR TITTLE

Queens College, City University of New York

MAX WEINER

Graduate School and University Center, City University of New York

FRED D. PHELPS

Lehman College, City University of New York

The present study was concerned with the validity of the College-Level Examination Program General Examinations (CLEP) in Mathematics and English Composition. The two-fold purpose of the study was to provide an estimate of the number of credit hours likely to be earned if students took CLEP as well as to examine the interrelationships among the two previously mentioned subtests from CLEP, an end-of-year achievement test in Mathematics, and a prior measure in English Composition. First-year students at a senior college of the City University of New York were recruited for an experimental administration of the CLEP, the final examination from the first year mathematics course, and a college-developed English placement essay. Students with high scores on the American College Testing Program Examination (ACT) and high school averages were predominant in the sample selected for testing. From the data, inferences were made that (1) the CLEP Mathematics test could be used to grant credit in mathematics, but that the current cutting score should be examined in view of standards used in the course; (2) there was little relationship between CLEP English Composition scores and present college placement procedures for first-year English; and (3) the number of students who could earn college credit by examination was much higher than was the number presently taking CLEP at the college.

THE use of examinations to grant credit for college level courses has

<sup>1</sup> The authors acknowledge with thanks the support of the College Entrance Examination Board in providing examinations and scoring services for this study.

been argued for on the basis of benefits to students (either in expanding or accelerating their educational experiences) and to institutions (in terms of allocation of staff and possible reduction of costs). The present report is the first one in a study of credit-by-examination at the City University of New York (CUNY), with one of the senior colleges agreeing to participate. Its dual purpose was to provide, for a sample of college freshmen, an estimate of the number of credit hours likely to be earned if students took the College-Level Examination Program General Examinations (CLEP) and to ascertain the degree of interrelationships of scores on each of two CLEP subtests (Mathematics and English Composition), of performance on an end-of-year achievement test in mathematics prepared by faculty members (but administered at the start of the fall semester), and of standing on a placement test (writing sample) in English administered three months prior to fall enrollment in freshman classes.

### *Methodology*

Letters to recruit volunteers to take CLEP were sent to 240 students. These letters were directed primarily to students in the upper part of the composite score distribution of the American College Testing Program Examination (ACT) and in the upper part of the distribution on high school average (HSA). Of this original group of 240 students, 181 agreed to participate; complete data for analyses were available for 171 students (69 males and 102 females). The ACT and HSA data for the 171 students were: ACT score range from 14 to 32, mean—23.8, standard deviation—3.46; HSA range from 73 to 96, mean—85.6, and standard deviation—5.61. After being tested September 8 and 9, 1973, students were granted college course credit on any CLEP test on which they had placed at or above cutting scores the college had established.

The tests administered were: CLEP General Examinations (5 scores in English Composition, Natural Sciences, Mathematics, Humanities, and Social Sciences—History; and a faculty constructed final examination for the college first year mathematics course, (FME). English writing samples were available from a June 1973 placement essay devised at the college for first-year English (placement in English 101 or 102); these essays were rescored in the fall, 1973, and given a grade of A, B, C or D by one of three raters.

The CLEP and FME were given on 2 days, with 2 testing orders randomly assigned (balanced for the FME): Group A took FME, then CLEP Books I and II; Group B took CLEP Book I (English Composition, Natural Sciences, Mathematics), FME, then CLEP Book II

(Humanities and Social Sciences—History). Examination of scores associated with test order (A,B) and date of testing (Saturday, Sunday) by analysis of variance revealed no significant main order or interaction effects for CLEP English Composition and Mathematics scores, or for the FME scores. These data from groups A and B and from two testing dates were combined for the remaining analyses.

### *Results and Discussion*

#### *Mathematics*

Correlations of the FME raw scores and grades (ranging from A to F) with CLEP Mathematics scores were .62 and .58, respectively. Correction for restriction in range, which was based on the use of ACT scores for the full first-year class of 1972 (mean of 16.5 and standard deviation of 5.35), resulted in only slightly higher correlation coefficients of .64 and .62, respectively. The recommended cutting score on CLEP Mathematics (495) was contrasted with grades of C or higher on the FME. This analysis showed that all those earning 0 credit on CLEP received a mark of D or lower on FME ( $N = 51$ ). There were no students who obtained a grade of C or higher on FME and also earned zero CLEP credit hours. Of those receiving 6 credit hours in mathematics on the basis of CLEP, 85 (71%) received a grade of D or lower on FME, and 35 (29%) earned a grade of C or higher.

Although the correlation of performance on the CLEP and FME was within the range of expected validity coefficients, the cutting score might need to be adjusted for the course standard. Examination of the scatterplot of FME grades and CLEP Mathematics showed that a grade of F was associated with a CLEP Mathematics score range from 361 to 604. Adjusting the cutting scores to reduce the proportion of those receiving CLEP credit and earning grades of D or lower would increase the numbers of students receiving C or higher grades who also earned zero CLEP credit. For example, cutting scores are shown in Table 1. Increasing the cutting score required for receiving credit could effect a considerable reduction in the numbers receiving grades of D or lower as compared with the frequency associated with the cutting score employed. The number of students who would have received a grade of C or higher on the basis of the FME but would not have received 6 credits on CLEP went from zero with the present cutting score of 495, to 3, 7, and 9 with use of higher cutting scores of 527, 543, and 559, respectively, on CLEP. Cutting scores would need to be established locally (and cross-validated) where grades on college final examinations for courses indicate a different standard from that established by use of CLEP recommended cutting scores.



TABLE I

*A Comparison of Number of Students with Zero or Six CLEP Credit Hours in Mathematics and with Faculty-Assigned Grades of C or Higher and D or Lower as CLEP Mathematics Cutting Scores Are Increased*

Number of students receiving zero CLEP credit hours			Number of students receiving six CLEP credit hours		
CLEP Mathematics Cutting Score	Number with grade of C or higher	Total number	CLEP Mathematics Cutting Score	Number with grade of D or lower	Total number
495 or higher	0 (0%)	51	495 or higher	85 (71%)	120
527 or higher	3 (4%)	83	527 or higher	56 (64%)	88
543 or higher	7 (7%)	98	543 or higher	45 (62%)	73
559 or higher	9 (8%)	115	559 or higher	30 (54%)	56

### *English*

The correlation between CLEP English Composition and grades on the English placement essay was .24 (.26, when corrected for restriction in range). This figure was consistent with the contingency coefficient of .22 computed for the four-fold table (receipt of either 0 or 6 hours CLEP credit vs. placement in either English 101 or English 102), the entries for which were based on readings of the essay the prior June (chi square = 8.35,  $p = .004$ ). These correlational results were lower than would ordinarily be useful for a study of the validity of granting credits based on examination.

### *Course Credits*

The number of students receiving 0 or 6 course credits for each of the 5 CLEP General Examinations was computed. Six hours of course credits were given students on the basis of cutting scores as follows: English Composition, 495; Natural Sciences, 485; Mathematics, 495; Humanities, 468; Social Sciences—History, 470. The percentage of students receiving credit ranged from 46 in English Composition to 70 in Mathematics; 51% received 6 hours of credit in Humanities, 53% in Social Sciences—History and 58% in Natural Sciences. Only 14 students (8%) of the 171 students failed to earn any course credit. Sixteen per cent (28 students) earned the maximum possible 30 credits, equivalent to a full year's credit. The rest of the distribution was: 29 students (17%) earned 6 hours credit; 26 (15%), 12 hours of credit; 42 (25%), 18 hours of credit; and 32 (19%), 24 credit hours by examination.

The number of students voluntarily taking CLEP prior to entering

the college (i.e., not recruited for this study) was 40. The results of the study clearly showed that many more students were capable of taking and receiving course credit on the basis of CLEP. The implications for the college enrollment were estimated by computing full time equivalent student (FTE) on the basis of credits earned by these 171 students. The total of 2850 credits (divided by 30 credits per student to obtain the number of FTE's), yielded the equivalent of 95 FTE students. Using a cost of \$1710 per FTE (instructional costs only), savings for 95 FTE students would be \$162,450. Projected for ten senior colleges of CUNY, savings would be \$1,624,500 annually. As gross estimates, these figures do not take into account a full cost analysis that would be required if large numbers of students were to earn credit by examination.

### *Conclusions*

The data reported in this investigation have indicated that studies of the validity of CLEP for granting course credit by examination should be carried out on an individual institution basis. Satisfactory validity was obtained for CLEP Mathematics, when a college course final examination in Mathematics was used as the criterion. However, the currently recommended cutting score might be too low, when examined against faculty standards. Similar validity was not demonstrated for CLEP English Composition, when a college English placement essay was employed as the criterion measure. The well-known difficulty in obtaining reader reliability for English essays undoubtedly contributed to the low correlations reported.

Comparison of the numbers of students voluntarily taking CLEP with the results of credits granted by examination in this study indicated that there would be a large potential group of students who could take the CLEP and earn college credit. This study did not attempt to anticipate long-range effects of encouraging credit by examination. Two of the many possible outcomes of granting college credit on the basis of CLEP scores would include shorter time in college and an alteration of the ratio of beginning to advanced courses within academic departments. These and other results should be examined in further testing of the validity of a program such as CLEP.



## PREDICTION OF FIRST QUARTER FRESHMAN GPA USING SAT SCORES AND HIGH SCHOOL GRADES

BRAD S. CHISSOM AND DORIS LANIER  
Georgia Southern College

The study attempted to determine the validity of students' SAT scores and HSGPA as predictors of freshman course grades and overall college grade point average (CGPA). Subjects for the study included 669 freshman students at Georgia Southern College who had enrolled in and completed either English Composition I or freshman mathematics during fall quarter, 1973. Data for the study included students' SAT-V scores, SAT-M scores, HSGPA and CGPA. Results showed that a significant multiple correlation existed between the predictor variables and CGPA.

MANY colleges and universities have adopted the College Entrance Examination Board's Scholastic Aptitude Test (SAT) as one criterion for selecting candidates for college admission. In many instances a student's SAT score, in combination with his high school grade point average (HSGPA), is the basis for his acceptance or rejection by the school of his choice. Since college admissions officers place a great deal of emphasis on a student's HSGPA and SAT score, continuous efforts should be made to determine the predictive validity of these scores.

Most researchers have found a correlation between first quarter or first year grades of college students and (a) high school grade point average (HSGPA) and (b) SAT scores. For example, Franz, Davis, and Garcia (1958) obtained a substantial correlation between first quarter grade averages of students and each of two cognitive predictors: HSGPA and SAT scores. Several investigators have concluded that HSA was a more valid predictor of college success than were SAT scores (Franz, Davis, and Garcia, 1958; Mann, 1961; Michael and Jones, 1963; Spaulding, 1959). Among the studies concerned with grade predictions in specific subject areas, Passons (1967) concluded

that, although high school achievement was the most predictive indicator of future over-all college success, test scores were slightly more valid than was high school standing for predicting grades in specific courses. In another investigation Brown and Lightsey (1970) obtained correlations of .54 and .50 for men and women, respectively, between SAT verbal (SAT-V) scores and English course grades. More recently, Lanier and Lightsey (1972) found a correlation of .74 between SAT-V scores and English grades and of .66 between HSGPA and college English grades. In conclusion, most correlational studies have revealed a substantial relationship between cognitive predictor variables and college achievement.

### *Purpose and Method*

The purpose of this study was to determine the validity of students' SAT scores and HSGPA as predictors of freshman course grades and overall college grade point average (CGPA) at Georgia Southern College. It was hypothesized that a positive correlation would exist between (1) SAT verbal scores (SAT-V) and CGPA, (2) SAT mathematics scores (SAT-M) and CGPA and (3) HSGPA and CGPA. The subjects for the study included all freshman students ( $N = 669$ ) who had enrolled in and completed either English Composition I or freshman mathematics during fall quarter, 1973, at Georgia Southern College. The following data were obtained from each student's record: (1) SAT-V scores, (2) SAT-M scores, (3) HSGPA, and (4) CGPA at the end of the fall quarter. Grade point average was based on a 4-point scale.

### *Results and Discussion*

Intercorrelations among the four variables along with means and standard deviations are included in Table 1. All three predictors correlated significantly with the CGPA criterion measure.

TABLE 1  
*Intercorrelations of Predictor and Criterion Variables Along with  
Their Means and Standard Deviations*

Variables	1	2	3	4	M	SD
1. SAT-V	—	.49*	.21*	.37*	421.55	76.05
2. SAT-M		—	.17*	.39*	458.01	82.78
3. HSA			—	.46*	2.81	.55
4. CGPA				—	2.35	.80

\* Significant at .01 level ( $R \geq .11$ ;  $p < .01$ ).



TABLE 2  
*Summary Table for Step-Wise Multiple Regression*

Variable Entered	Multiple		Increase in $R^2$
	$R$	$R^2$	
1. HSGPA	.45	.20	.20
2. SAT-M	.55	.30	.10
3. SAT-V	.57*	.32	.02

\*  $F = 108.244$  ( $df = 3,665$ ),  $p < .01$ .

Results of the step-wise multiple regression analysis are presented in Table 2. The numbers of the variables indicate the order in which they entered the prediction equation. The overall multiple correlation coefficient of .57, which was statistically significant beyond the .01 level, was comparable with coefficients obtained from other studies using the same three types of predictor variables. The largest contribution to the relationship was made by HSGPA followed by the SAT-M and SAT-V variables. It would appear that SAT-M was weighted more heavily for predicting CGPA for this group of subjects than was the SAT-V, but that the weighted combination of the three variables indicated limited validity for prediction of CGPA from either the SAT-V or the SAT-M. It was evident that HSGPA was the most valid predictor.

## REFERENCES

- Brown, J. L. and Lightsey, R. Differential predictive validity of SAT scores for freshman college English. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1970, 30, 961-65.
- Franz, G., Davis, J. A., and Garcia, D. Prediction of grades from pre-admissions indices in Georgia tax-supported colleges. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1958, 18, 841-42.
- Lanier, D. and Lightsey, R. Verbal SAT scores and high school averages as predictors. *Intellect*, 1972, 101, 127-28.
- Mann, Sister M. J. The prediction of achievement in a liberal arts college. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1961, 21, 481-83.
- Michael, W. B. and Jones, R. Stability of predictive validities of high school grades and of scores on the Scholastic Aptitude Test of the College Entrance Examination Board for liberal arts students. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1963, 23, 375-78.
- Passons, W. R. Predictive validities of the ACT, SAT, and high school grades for first semester GPA and freshman courses. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1967, 27, 1143-44.
- Spaulding, Helen. The prediction of first-year grade averages in a private junior college. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1959, 19, 627-28.



## THE RELATIONSHIP BETWEEN ACADEMIC APTITUDE AND OCCUPATIONAL SUCCESS FOR A SAMPLE OF UNIVERSITY GRADUATES

JOHN LEWIS

Winona State College

The academic aptitude at the time of admission to college and level of occupation later in life were collected for 619 male college graduates. The statistically significant relationship indicated that the graduates with higher aptitude scores as compared with those with lower scores were more likely to report higher level occupations.

THIS study was undertaken to determine whether academic aptitude at the time of admission to college was related to occupational success later in life among 619 male graduates.

### *Procedures*

Each subject's academic aptitude was defined as his composite score on the Iowa Placement Tests. Designed to predict academic achievement in college, this battery of tests was given to all students who entered the University of Iowa during the years prior to the development of the American College Testing Program Examinations. Occupational level was determined by the subjects' responses to questionnaires in the late 1960's that were mailed by the author and the University of Iowa Alumni Office. Roe's (1956) classification scheme was used to determine the level of each graduate's occupation. The subjects were 619 male graduates of the University of Iowa College of Liberal Arts in the academic years 1948-49, 1954-55, and 1959-60 who had majored in the areas of general humanities, social science, natural science, or journalism and for whom complete data could be found.

### *Findings*

The classification of the graduates by academic aptitude and later

TABLE I

*Ranges of Scores in Terms of Percentile Ranks on Freshman Test Battery and Corresponding Occupational Levels by Percentages of Individuals within These Test Score Ranges*

Occupational Level	Percentile Ranks in Test Scores			
	01-49	50-74	75-89	90-99
1	9	16	17	24
2	72	74	72	66
3	19	10	11	10
<i>N</i>	185	158	143	133
$\chi^2 = 20.57$	<i>df</i> = 6	$p < .01$		

occupational level is presented in Table I. Larger percentages of the graduates who had earned higher scores on the admissions tests as compared with those who had obtained lower scores reported occupations with level 1 classification. Likewise, larger percentages of graduates with low scores as compared with those who had received high scores reported level 3 occupations. For example, 9% of the graduates with admission test scores from the first to the forty-ninth percentile range reported level 1 occupations whereas 24% of those within the ninety to the ninety ninth percentile range on the tests reported level 1 occupations. The chi square value of 20.57 was statistically significant beyond the .01 level.

### Conclusion

These results show that the ability to achieve high scores on a typical college admissions test was related to occupational success among those graduates who responded to the questionnaire when occupational success was defined as the prestige level of their occupation.

### REFERENCE

- Roe, Anne. *The psychology of occupations*. New York: Wiley, 1956, 148.

## PREDICTING ACHIEVEMENT IN AN UPPER-DIVISION BACHELOR'S DEGREE NURSING MAJOR

JOHN LEWIS AND MARGARET WELCH  
Winona State College

A study of the correlations between objective background variables and achievement in an upper-division bachelors degree program in nursing revealed significant correlations for grade point average in required college pre-nursing courses, grade point average in elective college pre-nursing courses, and rank in high school graduating class. The results of a multiple regression analysis showed that grade point average in required pre-nursing courses was the only variable to yield a significant regression weight.

THE purpose of the study was to determine (1) what the degree of relationship between selected objective background variables and academic achievement in an upper-division bachelors degree program in nursing would be (2) which of these background variables would add to the efficiency of a regression function for predicting achievement in this nursing curricula, and (3) what the accuracy of this optimum regression function would be.

### *Procedures*

The subjects were 104 juniors and seniors in the bachelors degree nursing program at Winona State College in Winona, Minnesota. The background variables selected for investigation were grade point average in required college pre-nursing courses taken prior to formal application to the nursing program, grade point average in elective college courses taken prior to formal application to the nursing program, composite standard score on the American College Testing Program Examinations (ACT), rank in high school graduating class, and total number of elective college credits. The criterion was grade point average in six core nursing courses typically taken as juniors.



TABLE 1

*Correlations among Grades in Required Pre-Nursing Courses (GPAR), Grades in Elective Pre-Nursing Courses (GPAE), Composite Standard Score on the ACT Tests (ACT), Rank in High School Graduating Class (HSR), and Number of Elective College Credits (NEC), and Grades in Criterion of Core Nursing Courses (GPAC) N = 104*

	GPAR	GPAE	ACT	HSR	NEC	GPAC
GPAR	1.00	.41**	.44**	.45**	.04	.44**
GPAE	.41**	1.00	.21*	.27**	-.19	.25*
ACT	.44**	.21*	1.00	.40**	.07	.12
HSR	.45**	.27**	.40**	1.00	-.12	.28**
NEC	.04	-.19	.07	-.12	1.00	.00
GPAC	.44**	.25*	.12	.28**	.00	1.00

\*  $p < .05$ .

\*\*  $p < .01$ .

### Results

The correlations among all variables are presented in Table 1 and the results of a multiple regression analysis are shown in Table 2. Three objective background variables, grade point average in required college pre-nursing courses, grade point average in elective college pre-nursing courses, and rank in high school graduating class yielded statistically significant correlations with grades in the upper division core nursing courses. The relatively large correlation of .44 between grades in the required pre-nursing courses and later success in the nursing program probably reflects a high degree of relevance between these required pre-nursing courses and the core courses in the nursing major. The remaining background variables of composite standard scores on the ACT tests and of total number of elective college credits did not correlate significantly with grades in the nursing major.

Reference to the multiple regression data present in Table 2 reveals that grade point average in required pre-nursing courses was the only background variable which yielded a statistically significant ( $p < .01$ ) regression weight. The other four predictor variables, which had non-

TABLE 2

*Raw Score Regression Weights (b), Standard Score Regression Weights (B) and Multiple Correlations (R) for Each of the Background Variables*

	b	B	R
GPA Required Pre-Nursing Courses	.425**	.409**	.44
High School Rank	.348	.125	.45
American College Tests	-.019	-.126	.46
GPA Elective Pre-Nursing Courses	.069	.077	.47
Number of Elective Courses	.001	.021	.47

\*\*  $p < .01$ .

significant ( $p > .05$ ) regression weights, added little to the accuracy of a regression function. The single-order correlation between grade point average in required pre-courses and grades in the nursing program was .44 compared with a multiple correlation of only .47 when all five of the variables were used as predictors.

### *Summary*

This study was undertaken to determine an optimum method of predicting academic achievement in an upper-division bachelors degree program in nursing. The findings of the study showed that grade point average in required pre-nursing courses, grade point average in elective pre-nursing courses, and rank in high school graduating class correlated significantly with a criterion of achievement in this nursing program. However, a multiple regression analysis of the data revealed that grade point average in required pre-nursing courses was the only variable to yield a significant regression weight. Optimum prediction could thus be made using this one background variable as a predictor.



## PREDICTIVE VALIDITY OF THE GRADUATE RECORD EXAMINATION AND THE MILLER ANALOGIES TESTS

JOHN L. NAGI  
Hudson Valley Community College

For a sample of 63 graduate students, 33 of whom did complete and 30 of whom did not complete a doctoral program in Educational Administration at the State University of New York at Albany, statistically nonsignificant point biserial coefficients of 0.140 and 0.087 were determined respectively for the total scores on the aptitude portion of the Graduate Records Examination and scores on the Miller Analogies Test relative to the dichotomous criterion of completion or lack of completion.

THIS study was intended to determine the validity of the total score on the Graduate Record Examinations Aptitude Test (GRE) and of the score on the Miller Analogies Test (MAT) as predictors of a criterion of completion or noncompletion of the Doctoral Program in Educational Administration at State University of New York at Albany.

### *Source of Data*

The data required to determine the validity of the GRE and MAT scores as predictors of program completion were gathered by examining the student records kept by the Graduate Admissions office of the School of Education of the State University of New York at Albany. This examination provided GRE and MAT scores for 33 students who had completed the Doctoral Program and 30 who had not.

### *Results and Discussion*

To determine the degrees of relationship between GRE scores and a criterion variable of completion or non-completion of the program

and the extent of association between MAT scores and the same criterion variable, point biserial coefficients of correlation were computed. The respective coefficients of 0.140 ( $N = 63$ ) and of 0.087 ( $N = 63$ ) for the GRE and MAT predictor variables failed to reach statistical significance at the .05 level.

Several prior studies have indicated that the usefulness of the MAT in predicting success in graduate school has not been noteworthy, (Gill and Marascuilo, 1967; Hall and Robertson, 1964; Hyman, 1957; and Platz, McClintock, and Katz, 1959). In a study at Utah State University, Borg (1963), determined that the GRE was of little value in predicting completion of a doctoral program and obtaining a degree. The present study appears to bear out earlier findings that the GRE and MAT are not substantially valid predictors of program completion.

### REFERENCES

- Borg, W. R. GRE aptitude scores as predictors of GPA for graduate students in education. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1963, 23, 379-382.
- Gill, G. and Marascuilo, L. A. Measurable differences between successful and unsuccessful doctoral students in education. *California Journal of Educational Research*, 1967, 18, 65-70.
- Hall, E. and Robertson, M. Predicting success in graduate study. *Journal of General Psychology*, 1964, 71, 359-365.
- Hyman, S. The Miller Analogies Test and the University of Pittsburgh Ph.D.'s in psychology. *American Psychologist*, 1957, 12, 35-36.
- Platz, A., McClintock, C., and Katz, D. Undergraduate grades and the Miller Analogies Test as predictors of graduate success. *American Psychologist*, 1959, 14, 285-289.



## THE PREDICTIVE VALIDITY OF THE WPPSI WITH ISRAELI CHILDREN<sup>1</sup>

AMIA LIEBLICH AND MAYA SHINAR

The Human Development Center  
The Hebrew University of Jerusalem, Israel

Sixty-two first grade children in Israel were tested on the WPPSI, and most of them were retested a year later using objective measures of Reading and Arithmetic. The scores of the intelligence and achievement tests were correlated to provide estimates of the predictive validity of the WPPSI. Results revealed the Israeli WPPSI to be highly valid for prediction of school achievement.

SINCE the publication of the WPPSI (Wechsler, 1967) several evaluation studies were carried out focussing on its relationship to other measures of intelligence (Yule, Berger, Butler, Newham, and Tizard, 1969; Zimmerman and Woo-sam, 1970) and on its predictive validity (Rankin and Henderson, 1973; Kaufman, 1973). The validity of the test as a predictor of early school achievement was rather satisfactory in the United States with middle-class children (Kaufman, 1973), but nonsignificant with disadvantaged Mexican-American children (Rankin and Henderson, 1973).

The purpose of the present study was to assess the predictive validity of the WPPSI in Israel, through use of a standardized Hebrew version of the test (Lieblich, 1971).

### *Method*

#### *Sample*

Sixty-two first grade children, their ages ranging from five and a half to six and a half years, participated in the first stage of the study. They

<sup>1</sup> The authors wish to thank Dr. David Wechsler for his constant help and support, and Mrs. M. Bassok for carrying out the individual examinations.

were randomly sampled from the first grade pupils in a public school in an urban middle-class area. Fifty-four of these children were re-located eighteen months later for the second stage of the project.

### *Instruments and Procedure*

In the first stage, the Hebrew WPPSI was administered individually during 1973. A year and a half later, 54 subjects who had been re-located in the second grade were given Israeli objective group tests of school achievements in reading and arithmetic (Ben Shachar and Ortar, 1968; Minkowich, 1973.)

### *Results and Discussion*

The WPPSI and the achievement tests were scored and WPPSI scores were converted to scaled scores through using the appropriate age norms. Correlations of the scaled scores of the subtests, the IQ scores of the Verbal, Performance and Total scales were computed separately with each of the two achievement criteria. The coefficients appear in Table 1.

The correlation between the reading and arithmetic tests was .67; between Verbal and Total IQs, .95; Performance and Total IQs, .91; and Verbal and Performance IQs, .74. That all the coefficients in Table 1 were statistically significant ( $p < .05$ ) indicated that even predictions of the criterion measures from the individual subtests were surprisingly valid, especially for the Arithmetic test.

TABLE 1  
*Correlations of WPPSI Scaled Scores of the Subtests and of Verbal Performance and Total IQs with Each of the Criterion Measures: Arithmetic and Reading Tests*

WPPSI subtests		Arithmetic	Reading
Verbal	Information	.59	.34
	Vocabulary	.43	.49
	Arithmetic	.54	.44
	Similarities	.54	.53
	Comprehension	.45	.40
Performance	Animal House	.41	.46
	Picture Completion	.45	.44
	Mazes	.61	.44
	Geometric Designs	.58	.51
	Blocks	.57	.26
Verbal IQ		.64	.57
Performance IQ		.73	.61
Total IQ		.73	.63

Apparently, the predictive ability of the WPPSI in Israel was found to be highly satisfactory. The reported correlations were higher than the ones found in American studies of the WPPSI, but resembled those found with some other mental tests for similar age groups (Kaufman and Kaufman, 1972).

To some degree this high predictability of the WPPSI could be hypothesized as being attributable to the heterogeneity of the sample, as the school underwent recent integration with a lower-class area. Additional validity statistics on the Israeli culture and subcultures are needed to determine whether these preliminary findings can be replicated.

### REFERENCES

- Ben Shachar N. and Ortar, G. Reading Comprehension Test, Israeli Ministry of Education, Jerusalem, 1968 (Hebrew).
- Kaufman, A. S. Comparison of the WPPSI, Stanford-Binet, and McCarthy scales as predictors of first-grade achievement. *Perceptual and Motor Skills*, 1973, 36, 67-73.
- Kaufman, A. S. and Kaufman, N. L. Tests built from Piaget's and Gesell's tasks as predictors of first-grade achievement. *Child Development*, 1972, 43, 521-535.
- Liebllich, A. *Manual for the Hebrew WPPSI*. Jerusalem, 1971 (Hebrew).
- Minkowich, A. *Arithmetic Test*. Jerusalem, Israel: The School of Education, The Hebrew University of Jerusalem, 1973 (Hebrew).
- Rankin, R. J. and Henderson, W. H. WPPSI reliability and predictive validity with disadvantaged Mexican-American Children. *Journal of School Psychology*, 1973, 11, 1.
- Wechsler, D. *WPPSI Manual*. The Psychological Corporation, N. Y., 1967.
- Yule, W., Berger, M., Butler V., Newham V., and Tizard J. The WPPSI: An empirical evaluation with a British sample. *The British Journal of Educational Psychology*, 1969, 39, 1-13.
- Zimmerman, I. L. and Woo-Sam J. The utility of the Wechsler Pre-school and primary scale of intelligence in the public school. *Journal of Clinical Psychology*, 1970, 26, 472.



## INTERRELATIONSHIPS AMONG PSYCHOLOGICAL MEASURES OF COGNITIVE STYLE AND FANTASY PREDISPOSITION IN A SAMPLE OF 100 CHILDREN IN THE FIFTH AND SIXTH GRADES

SUSAN McNARY, WILLIAM B. MICHAEL, LEO RICHARDS,  
AND CONSTANCE LOVELL

University of Southern California

For a sample of 100 fifth and sixth grade pupils of middle-class background (51 girls and 49 boys), intercorrelations among measures of three cognitive style constructs of reflection-impulsivity, field dependence-independence, and internal-external locus of control and three measures of fantasy predisposition were for the most part low and statistically not significant. Exceptions were noted in the instance of the sample of girls for whom coefficients of .36, .38, and .51 between the measure of locus of control and each of three measures of fantasy predisposition were statistically reliable beyond the .01 level. The hypothesis of a positive relationship between a measurable construct of fantasy predisposition and each of three measurable constructs of cognitive style received only limited support. Furthermore, it did not appear that fantasy was a part of a larger or more general construct of cognitive style or that a universal construct of cognitive style existed.

DURING the past several years considerable attention has been given to the study of fantasy, daydreaming, and imagination in children as being potentially facilitating to their cognitive and emotional development (e.g., Singer, 1973). Recently, Singer (1973, p. 220) has stressed the need to obtain evidence regarding whether a tendency toward make-believe play might not be but one component of a more general cognitive style such as reflection-impulsivity (Kagan, Rosman, Day, Albert, and Phillips, 1964) or field dependence-independence (Witkin, Oltman, Raskin, and Karp, 1971). In fact, another extensively investigated construct of locus of internal-external control, which was



derived from social learning theory (Rotter, 1954) and measured by Bialer (1961), would also appear, at least on the surface, to reflect many of the same activities suggested by the construct of field independence-dependence as well as an inclination toward autonomous or self-directed behavior patterns such as those involved in fantasy.

### *Problem*

Thus, the purpose of this correlational study was to determine for a group of 100 fifth- and sixth-grade children of middle class background the extent of the relationship if any between each of three measures of cognitive style (reflection-impulsivity, field dependence-independence, and internal-external locus of control) and each of two separate measures and a combined measure of fantasy predisposition. On the assumption that in their operational representation, reflection, field independence, and internal locus of control constitute positive poles, whereas impulsivity, field dependence, and external locus of control portray negative poles and on the further assumption that high standing on a measure of fantasy reflects its presence, the following two hypotheses within the context of the theoretical and empirical contributions of the previously cited researchers were suggested:

1. A positive relationship would occur between a measurable construct of fantasy predisposition and each of three measurable constructs of cognitive style.

2. Positive interrelationships would exist between pairs of measurable constructs of cognitive style.

High positive interrelationships among all four constructs cited might suggest that each was a subcomponent of a relatively general or ubiquitous construct of cognitive style. A study of the pattern of correlations among measures would also furnish some information regarding the extent to which measures of the selected constructs were independent of or related to one another as well as predictive of one another. Some inferences about construct validity might also be possible. Since many investigators (Singer, 1973; Pulaski, 1970; and Torrance, 1962) have considered imagination and fantasy-related activities as cognitive skills that reflect many of the same abilities such as flexibility, ideational fluency, and originality found in creative endeavors, the demonstration of significant relationships between measures of the cognitive styles cited and those of fantasy predisposition might suggest important implications for nurturing creative problem-solving skills in children.

## *Methodology*

### *Sample*

Residing in a middle-class suburban community near Los Angeles, the 100 fifth- and sixth-grade children (51 girls and 49 boys) varied in age from 10 years, 7 months to 13 years, 2 months.

### *Instrumentation*

To represent the constructs of reflection-impulsivity, field dependence-independence, and internal-external locus of control, respectively, the measures or scales entitled Matching Familiar Figures (MFF) yielding separate time scores and error scores (Witkin, et al., 1971), Embedded Figures Test (EFT) furnishing a time score only (Kagan, et al., 1964), and Children's Locus of Control Scale (CLOCS) providing a score of the number of correctly answered items, were employed. On the MFF and EFT high time scores were considered to indicate, respectively, reflection and field dependence, whereas low time scores suggested, respectively, impulsivity and field independence. On the MFF a high error score was interpreted as revealing impulsivity, and a low error score, reflection.

The construct of fantasy predisposition was defined operationally as the (a) *M* score derived from the Holtzman Inkblot Technique (HIT) (Holtzman, 1963), (b) the total score from Forms A and B of the Torrance Tests of Creative Thinking (TTCT)—Thinking Creativity with Words, Activity 7: Just Suppose (Torrance, 1966, 1974), and (c) an unweighted sum of scores (a composite score) on these two measures. Although it could be argued rather convincingly that a composite score might reveal a complex construct reflecting quite varied psychological processes, the correlational analysis of each individual fantasy measure which contributed to the composite score as well as of the composite score with each of the other four measures of cognitive style would prevent any important information loss.

In addition to the seven psychological variables just enumerated, a measure of chronological age was also included. All seven test variables, their intended constructs, and the age variable are cited in Table 1.

### *Test Administration*

The first cited author and a trained assistant administered the five tests employed in a specific order during a 4-day period at each of four

TABLE 1  
*Intercorrelations of Measures Representing Constructs of Cognitive Style and Predisposition to Fantasy and of Age Along with Ranges, Means, and Standard Deviations of Scores on Each Variable<sup>a,b</sup>*

Test Variables and Age Variable	Group or Sub-group	Intended Construct	Correlation Coefficients								Descriptive Statistics					
			(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	Range	Mean	SD			
(1) Matching Familiar Figures—Time Score	Total	Reflection-Impulsivity	—	—	22	19	05	—	04	—	03	—	04	2.9-38.5	12.63	5.74
	Boys		—	—	21	14	19	00	23	18	07	—	07	2.9-38.5	13.60	6.26
	Girls		—	—	21	31	07	04	09	09	14	—	14	5.1-29.9	11.62	5.14
(2) Matching Familiar Figures—Number of Errors Score	Total	Impulsivity-Reflection	—	—	—	30	08	04	—	04	—	01	—	0-22	9.95	4.87
	Boys		—	—	—	35	14	—	02	26	20	—	07	0-22	9.61	5.55
	Girls		—	—	—	22	01	09	10	13	05	—	05	1-19	10.27	4.14
(3) Embedded Figures Test—Time Score	Total	Field Dependence—Independence	19	30	—	—	06	09	08	02	—	14	—	12.6-180.0	85.60	35.90
	Boys		14	35	—	—	11	05	23	20	00	—	00	12.6-180.0	80.81	36.18
	Girls		31	22	—	—	25	28	11	24	29	—	29	33.2-171.2	90.21	35.36
(4) Children's Locus of Control Scale	Total	Internal-External	05	08	—	06	—	21	23	29	11	—	11	6-21	14.78	2.75
	Boys		19	14	11	—	03	00	02	06	—	06	—	6-20	14.55	2.82
	Girls		07	01	25	—	36	38	31	16	—	16	—	7-21	15.00	2.69
(5) Torrance Tests of Creative Thinking—Thinking Creatively with Words, Activity 7; Just Suppose	Total	Fantasy	—	04	09	21	—	11	56	—	04	—	04	0-39	14.20	7.72
	Boys		00	02	05	03	—	14	67	02	—	02	—	0-37	12.82	7.77
	Girls		04	09	28	36	—	02	45	—	11	—	11	2-39	15.53	7.50
(6) Holtzman Inkblot Technique (Group Test)	Total	Fantasy	—	01	04	08	23	11	—	88	15	—	15	1-62	21.91	13.48
	Boys		23	26	23	00	14	—	83	00	—	00	—	1-49	16.73	10.48
	Girls		09	10	11	38	02	—	88	25	—	25	—	6-62	26.88	14.24
(7) Unweighted Sum of Variables (5) and (6)—Total Fantasy Score	Total	Fantasy	—	03	01	02	29	56	88	—	11	—	11	10-84	36.11	16.24
	Boys		18	20	20	02	67	83	—	01	—	01	—	10-61	29.55	13.90
	Girls		09	13	24	51	45	88	—	17	—	17	—	15-84	42.41	15.93
(8) Age	Total	Fantasy	—	04	01	14	11	04	15	11	—	—	—	127-158	139.95	7.12
	Boys		07	07	00	06	02	00	01	—	—	—	—	129-146	139.65	6.96
	Girls		14	05	29	16	11	25	17	—	—	—	—	127-158	140.24	7.34

<sup>a</sup> Coefficients of absolute magnitudes of 17, 24, and 24 for the total sample, sample of boys, and sample of girls, respectively, required for statistical significance at the .05 level for

schools. Working in separate rooms, the examiners gave individual administrations of the MFF, EFT, and the TTCT subtest. The CLOCS and HIT were administered to groups of approximately 10 children approximately one or two days after the examinees had completed the first three tests.

### *Data Analysis*

Intercorrelations of scores on the seven test variables and chronological age were effected for the total sample as well as for separate subsamples of 49 boys and 51 girls. For each of these three groups, ranges, means, and standard deviations of scores on each variable were calculated and are reported in Table 1. In view of the directional nature of the research hypotheses one-tailed significance tests were employed.

### *Findings*

In terms of the correlational entries in Table 1, the following results may be summarized.

1. Although for the total group the correlation coefficients between scores on scales of reflection-impulsivity and of field dependence-independence and scores on each of the three fantasy measures were small (varying from  $-.09$  to  $.08$ ) and statistically not significant, statistically reliable but low positive correlation coefficients ( $.21$ ,  $.23$ , and  $.29$ ) were found between a measure of internal locus of control and each of three measures of fantasy predisposition.

2. For the subgroup of boys, low correlation coefficients ranging from  $-.26$  to  $.23$  (only one being significant) between measures of fantasy predisposition and those of cognitive style were obtained.

3. Although for the subgroup of girls coefficients of  $-.28$  and  $-.24$  between a measure of field dependence and each of two measures of fantasy predisposition were statistically significant at the  $.05$  level, coefficients of  $.36$ ,  $.38$ , and  $.51$  between the measure of locus of internal control and each of the three measures of fantasy predisposition were statistically reliable beyond the  $.01$  level.

4. For all samples, the values of the intercorrelations among the measures of cognitive style fell between  $-.25$  and  $.35$  with only six coefficients attaining statistical significance.

5. With respect to the HIT measure (Variable 6) and the Total Fantasy Score (Variable 7) the mean of the subsample of girls was significantly higher than that of the boys ( $p < .01$ ).

6. In the instance of the subsample of girls, statistically significant

correlation coefficients of  $-.29$  and  $.25$ , respectively, occurred between age and (time) scores on the measure of field dependence and between age and scores on the HIT measure of fantasy predisposition.

### *Conclusions*

On the basis of the data obtained, the following conclusions were drawn:

1. In the main, only limited support was obtained for the first hypothesis and practically no support for the second hypothesis.
2. In light of the modest correlations obtained for girls between locus of internal control and fantasy predisposition it was suggested that girls entering or about to enter adolescence who show somewhat greater internal control might be expected to exhibit a higher level of predisposition to fantasy.
3. Fantasy would not seem to be part of larger or more general construct of cognitive style.
4. The presence of a universal or general construct of cognitive style would appear to be highly improbable.

### *Discussion*

That more striking evidence for the support of the two major hypotheses was not found could be attributed to a number of factors, particularly the unreliability of the measures employed. Time constraints imposed by the school program prevented the realization of test-retest estimates of reliability. Furthermore, the lack of homogeneity of items within many of the scales not only precluded the determination of interpretable estimates of internal consistency but also prevented the realization of estimates of potential maximum correlations that could have been attained between two sets of scores theoretically free of error variance.

One disconcerting finding was the pattern of positive and negative signs accompanying the three coefficients arising from the intercorrelations of the first three measures or variables cited in Table 1. Although the negative coefficients of  $-.22$ ,  $-.21$ , and  $-.21$  for the total sample, the subsample of boys, and the subsample of girls, respectively, between time scores of the MFF (Variable 1) and error scores of the MFF (Variable 2) were anticipated, as were the corresponding, positive coefficients of  $.30$ ,  $.35$ , and  $.22$  between Variable 2 (MFF—Error Score) and Variable 3 (EFT—Time Score), the *positive* coeffi-



cients of .19, .14, and .31 between Variable 1 (MFF—Time Score) and Variable 3 (EFT—Time Score) were expected to be negative. In other words, as expected, the correlations between measures of reflection (Variable 1) and impulsivity (Variable 2) were negative and those between impulsivity (Variable 2) and field dependence (Variable 3) were positive. Not anticipated, however, was the *positive* correlation between the measure of reflection (Variable 1) and that of field dependence (Variable 3), although admittedly only the positive correlation of .31 for the subsample of girls was statistically reliable. The examiners did note during the testing sessions a tendency on the part of a large proportion of children who displayed behavior interpreted as being field dependent and impulsive to perseverate in responses revealing errors or mistakes in performance. Hence, the extension in time associated with perseveration that took place in the completion of the MFF and EFT tasks (both scored for the amount of time required) might have accounted in part for the small positive correlation between reflection (Variable 1) and field dependence (Variable 2). Moreover, there was the distinct possibility that the unfamiliarity of the tasks and the lack of sufficient prior time spent on practice questions or in clarification of task requirements could have contributed a disproportionate increment to the time score—time that had to be spent in comprehending the nature of the task and in commission of several errors before the child could eventually figure out what he or she had to do. Thus in addition to a perseverative error, the additional time required of a relatively immature group to comprehend the nature of task requirements would suggest that the correlation reflected a time component artifact that in reality was unrelated to the intended perceptual and affective characteristics underlying the reflection-impulsivity and field dependence-independence constructs as conceptualized by theorists who defined them.

The existence of sex differences in the manifestation of fantasy could be accounted for in terms of a number of speculative factors such as the nature of the particular measures chosen, level of cognitive and perceptual maturation of the subsamples of boys and girls, richness of varied stimuli in the home environment permitting a basis for imaginative thinking (as related in turn to the socioeconomic status of the families), and the social expectations of significant others in the adult or peer culture regarding the desirability or acceptability of displaying fantasy-related behaviors. Although Biblow (1970) and Pulaski (1970) failed to find significant sex differences in their respective investigations of fifth-grade and kindergarten children, it is conceivable that differences observed in the two subsamples studied were related to a combination of several of the factors just cited.

### Recommendations

In view of the somewhat inconclusive results of this investigation, it is suggested that a need exists to examine within a comprehensive theoretical framework a number of constructs pertaining to cognitive styles, predisposition to fantasy, perseveration, creative abilities, and problem-solving skills as well as to develop appropriate instruments for their operational description. In combination with a sound and inventive conceptualization of affective and cognitive constructs use of the paradigm afforded by the multitrait-multimethod matrix devised by Campbell and Fiske (1959) would afford a useful methodology for clarification of the nature of the constructs studied.

Initiation of longitudinal studies would provide evidence concerning the influence of age, sex, socioeconomic background, and subcultures on the manifestation of fantasy and identifiable cognitive styles. It could be interpreted that applications of the Campbell and Fiske methodology on a longitudinal basis would do much to improve not only the clarity and meaningfulness of the constructs studied but also the validity and reliability of their empirical representation by a variety of tests and scales.

### REFERENCES

- Bialer, I. Conceptualization of success and failure in mentally retarded and normal children. *Journal of Personality*, 1961, 29, 303-320.
- Biblow, E. The role of fantasy in the reduction of aggression. *Dissertation Abstracts*, 1970, 31, 3699.
- Campbell, D. T. and Fiske, D. W. Convergent and discriminant validation of the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, 56, 81-105.
- Holtzman, W. Holtzman inkblot technique: Group administration by slide projection. *Journal of Clinical Psychology*, 1963, 19, 433-454.
- Kagan, J., Rosman, B., Day, D., Albert, J., and Phillips, W. Information processing in the child: Significance of analytic and reflective attitudes. *Psychological Monographs*, 1964, 78(1, Whole No. 578).
- Pulaski, M. A. Play as a function of toy structure and fantasy predisposition. *Child Development*, 1970, 41, 531-537.
- Rotter, J. B. *Social learning and clinical psychology*. Englewood Cliffs, N. J.: Prentice-Hall, 1954.
- Singer, J. L. *The child's world of make-believe: Experimental studies in imaginative play*. New York: Academic Press, 1973.
- Torrance, E. P. *Guiding creative talent*. Englewood Cliffs, N. J.: Prentice-Hall, 1962.
- Torrance, E. P. *Torrance tests of creative thinking: Thinking creatively with words—activity 7: Just suppose*. Princeton, N. J.: Personnel Press, 1966.

- Torrance, E. P. *Torrance Tests of Creative Thinking: Norms—Technical manual*. Lexington, Mass.: Personnel Press/Ginn and Company, Xerox Education Company, 1974.
- Witkin, H. A., Oltman, P. K., Raskin, E., and Karp, S. A. *A Manual for the Embedded Figures Tests*. Palo Alto, Calif.: Consulting Psychologists Press, 1971.



## A COMPARISON OF THE WRAT AND THE PIAT WITH LEARNING DISABILITY CHILDREN

DALE D. BAUM  
New Mexico State University

This study was concerned with the comparative performance of learning disability children on the WRAT and the PIAT. Correlations between corresponding subtests of the two instruments were quite high for Reading and moderately high for Spelling and Arithmetic across four age levels (7 to 8, 9, 10, and 11 year olds). It was concluded that the utility of the PIAT is as promising as that of the WRAT, although diagnosticians using the PIAT may wish to exclude the Reading Comprehension subtest, as this subtest tends to be measuring the same skills as Reading Recognition at the lower grade levels.

PSYCHOLOGISTS and educational diagnosticians responsible for the assessment of children with learning problems generally include a test of school achievement in their battery of instruments if current information is not readily available from teachers or from the school files. Typically, the instrument of choice has been an individually administered test of the screening or wide-range variety which would yield an overview of an individual's academic status.

The Wide Range Achievement Test (WRAT) in its various editions (Jastak, 1936, 1946; Jastak and Jastak, 1965) has enjoyed wide popularity as a relatively quick test of the basic school subjects of Reading, Spelling, and Arithmetic. Until the recent introduction of the Peabody Individual Achievement Test (PIAT) by Dunn and Markwardt (1970) the WRAT had maintained a virtually unchallenged position in educational diagnosis. Proger (1970) has suggested that the PIAT represents a sophisticated and formidable challenge to the WRAT.

In addition to the subtests Reading Recognition, Spelling, and



Mathematics, the PIAT also includes subtests of Reading Comprehension and General Information which have no counterpart on the WRAT. The PIAT also provides normative data for a total test score while the WRAT does not. The PIAT employs multiple choice selection for Reading Comprehension, Spelling, and Mathematics, while the WRAT measures Reading by word recognition, Spelling from dictation, and Arithmetic from computation. The WRAT reportedly requires 20 to 30 minutes to administer, whereas the PIAT requires 30 to 40 minutes.

In previous research by Sittington (1970) with educable mentally retarded adolescents and by Soethe (1972) employing normal, reading disabled, and mentally retarded children, the conclusion was reached that those subtests of the PIAT having counterparts on the WRAT demonstrated reasonably high concurrent validity. The present investigation was concerned with the comparative performance of learning disabled children of various ages on the WRAT and the PIAT.

### *Method*

The 82 males and 18 females who served as subjects for this study were randomly selected from self-contained classes for learning disabled students. The classes were located in primarily middle and upper-middle class suburban areas adjacent to a large southern city. Twenty-five subjects were selected from each of the four age groups: 7 to 8, 9, 10, and 11 year olds. The average age for the total group was 9 years, 7 months with a range from 7 years, 2 months to 11 years, 11 months.

Both the WRAT and the PIAT were individually administered to each subject in a single session. The order for presenting the two tests was reversed for every other subject in order to counterbalance any practice effects which might have accrued.

### *Results*

The means and standard deviations of grade equivalency scores for each of the subtests of the WRAT and the PIAT are presented in Table 1 for subjects at each of the four age levels studied. In general, the subtest scores of the WRAT did correspond more closely with their counterparts on the PIAT at the 7 to 8 year level than they did at the 11 year level. This finding, however, tends merely to reflect known characteristics of both tests, i.e., as age increases the standard error of measurement increases on both instruments.

Pearson product moment correlation coefficients between cor-

TABLE 1  
Means and Standard Deviations of Grade Equivalent Scores of the PIAT and WRAT by Age Level

Age	N	WRAT			PIAT					Total
		Spelling	Arith.	Reading	Spelling	Math	Reading Recognition	Reading Comprehension	Info.	
7-8	25	1.5	1.9	1.5	1.7	1.5	1.6	1.2	1.3	1.3
		.47	.72	.65	.80	.72	.67	.12	.12	.78
9	25	1.9	2.4	2.0	2.0	2.4	1.9	1.7	2.4	1.9
		.51	.44	.59	.72	.78	.67	.12	.13	.76
10	25	2.5	3.0	2.8	2.7	3.4	2.9	2.6	3.6	2.9
		.64	.66	.93	.68	.10	.11	.99	.11	.66
11	25	2.8	3.6	3.0	3.4	4.0	3.1	3.2	4.4	3.5
		.68	.74	.99	.17	.15	.11	.84	.14	.93

responding subtests of the WRAT and the PIAT are presented in Table 2. The WRAT Reading subtest was highly correlated (.85 to .89) with the PIAT Reading Recognition subtest across all four age levels. The PIAT Reading Comprehension subtest correlated quite highly (.90) with WRAT Reading at the 11 year level, moderately (.72 and .62) at the 7 to 8 and 9 year levels, and relatively low (.56) at the 10 year level. Correlations between the PIAT and WRAT Spelling subtests were moderate across all four age levels (.61 to .71), whereas the correlations between PIAT Mathematics and WRAT Arithmetic ranged from moderate (.77) at the 7 to 8 year level to quite low (.49) at the 11 year level.

The intercorrelations of the subtests of the two instruments are presented in Table 3. Employing the entire sample of 100 subjects, WRAT Arithmetic correlated .77 with PIAT Mathematics, whereas WRAT Spelling correlated .81 with PIAT Reading Recognition and .77 with PIAT Reading Comprehension. Of additional interest is the finding that the subtests of the WRAT did correlate as highly with the PIAT information and Total Test as did the remaining subtests of the PIAT. It should be noted, of course, that the heterogeneity in age of the total sample contributed to the realization of what must be viewed as relatively high coefficients.

### Discussion

In this study PIAT Reading Comprehension and Reading Recognition correlated .78 as compared with .84 in both the Sittington (1970) and Soethe (1972) studies. Since the first 18 items of the Reading Recognition subtest count as the first 18 items of the Reading Comprehension subtest, relatively high correlations between these

TABLE 2  
*Correlations between Selected Subtests of the PIAT and WRAT by Age Level*

Age Level N	7 to 8 25	9 25	10 25	11 25
PIAT				
Reading Recognition	.85**	.87**	.89*	.87**
Reading Comprehension	.72**	.62**	.56*	.90**
PIAT				
Spelling	.62**	.71**	.62**	.61**
PIAT				
Mathematics	.77**	.79**	.65**	.49*

\* Significant at .05 level.

\*\* Significant at .01 level.

TABLE 3  
*Correlation Matrix of the Subtests of the WRAT and PIAT<sup>a</sup> (N = 100)*

	WRAT				PIAT			
	Spelling	Arithmetic	Reading	Math	Reading Recog- nition	Reading Compre- hension	Spelling	Information
WRAT								
Spelling	—							
Arithmetic		73	81	67	81	77	73	68
Reading		—	61	77	57	62	56	68
			—	62	92	77	69	63
PIAT								
Mathematics				—	62	61	48	73
Reading					—	78	66	67
Recognition						—	67	66
Comprehension							—	54
Spelling								—
Information								
Total								

<sup>a</sup> All decimal points omitted from the correlation coefficients, all of which were significant beyond the 0.1 level.

subtests would be expected with subjects who function at the lower grade levels because of the overlapping content. Since the content validity of the Reading Comprehension subtest at the lower grade levels is questionable because of its sharing of the same content with Reading Recognition, this subtest should be interpreted cautiously by diagnosticians who employ the instrument with subjects functioning at the lower grade levels.

It was also observed that PIAT Reading Comprehension correlated .77 with both the WRAT Spelling and Reading subtests even though the first subtest is allegedly measuring comprehension; the second, spelling by dictation; and the third, word recognition. Similarly, PIAT Reading Recognition correlated .81 with WRAT Spelling and .92 with WRAT Reading. These relatively high intercorrelations suggested that a set of prerequisite skills peculiar to the language arts area operates fairly evenly across these subtests at the lower grade levels. Therefore, the language arts oriented subtests both within and between the PIAT and the WRAT share a commonality of both general educational content and psychological process which tends to lessen the specific content validity of these individual subtests at the lower grade levels.

### *Conclusion*

In view of the high degree of correlation between the two instruments, the findings of this study suggested that the utility of the PIAT as a wide-range achievement test is at least as promising as the WRAT has been. However, there appears to be little advantage in administering both PIAT Reading Comprehension and Reading Recognition to subjects functioning at the lower grade levels, since both subtests tend to be measuring the same skills.

When either of the two instruments is considered as a criterion, a substantial degree of concurrent validity relative to school achievement would appear to exist between like subtests of the two instruments. In view of the more comprehensive and diagnostic data possible with an additional two subtests and a total test score, it would appear that the PIAT can be employed confidently in the assessment of children with learning problems.

### REFERENCES

- Dunn, L. M. and Markwardt, F. C. Peabody Individual Achievement Test: Manual. Circle Pines, Minnesota: American Guidance Association, 1970.
- Jastak, J. F. Wide Range Achievement Test: Manual. Wilmington, Delaware: C. L. Story Co., 1936.



- Jastak, J. F. *Wide Range Achievement Test: Manual*. Wilmington, Delaware: C. L. Story Co., 1946.
- Jastak, J. F. and Jastak, S. R. *Wide Range Achievement Test: Manual*. Wilmington, Delaware: Guidance Associates, 1965.
- Proger, B. B. Test review number 4, Peabody Individual Achievement Test. *Journal of Special Education*, 1970, 4, 461-467.
- Sittington, P. L. *Validity of the Peabody Individual Achievement Test with educable mentally retarded adolescents*. Unpublished masters thesis. Honolulu: University of Hawaii, 1970.
- Soethe, J. W. Concurrent validity of the Peabody Individual Intelligence Test. *Journal of Learning Disabilities*, 1972, 5, 47-49.



## VALIDITY AND RELIABILITY OF A SIMPLE DEVICE FOR READINESS SCREENING

MARJORIE HAYES<sup>1</sup>  
CRESTWOOD KINDERGARTEN

EMANUEL MASON  
University of Kentucky

ROBERT COVERT  
University of Virginia

The utility of the Hayes Early Identification Listening Response Test (HEILRT), a rapidly administered screening test for readiness for first grade, was studied with 121 kindergarten pupils who were tested at the beginning of the academic year. The test could be administered in 15 to 20 minutes to groups of up to 30 children at a time. Reliability of the test was estimated to be .86 and rater reliability was .99. The test correlated highly positively (.79) with the Metropolitan Readiness Test, but not with age, number of siblings, or educational level of the mother. It was concluded that based upon the data, the HEILRT held promise as a readiness screening instrument for use with kindergarten children.

DECIDING the readiness of young children to begin work in school has long been an issue concerning educators. Standardized tests developed to assess the various facets of school readiness have tended to be quite long and arduous to administer to young children who are not used to testing and school routine (e.g., the Metropolitan Readiness Test and the Stanford Early School Achievement Test). Readiness testing could be more efficient if a reliable and valid initial screening instrument were available which broadly assessed readiness

<sup>1</sup> Copies of the materials developed for the Hayes Early Identification Listening Test can be obtained by writing to Mrs. Marjorie Hayes, Director, Crestwood Kindergarten, 1006 East Main Street, Frankfort, Kentucky 40601.

skills, and which took little time to administer and score. The purpose of the present study was to investigate the characteristics and utility of such a test, the Hayes Early Identification Listening Response Test (HEILRT). The HEILRT was designed to emphasize the importance of listening comprehension, visual perceptual, and fine motor skills. The test was developed over several years to screen young children for readiness rapidly and accurately by using tasks of the type with which children beginning school were usually familiar.

### *Method*

The HEILRT was administered to 121 beginning kindergarten pupils in a private school in central Kentucky. The mean age of the group at the time of the test administration was 62 months. The test was administered by the staff of the school to groups of 15 to 30 children at a time during the first month of school.

The HEILRT contains a series of psychomotor tasks for which the teacher gives verbal instructions. Each child is initially given a piece of blank paper and a box of coloring crayons. Standardized instructions require that the teacher show several examples at the blackboard. Then tasks are given children to perform at their tables, such as "Draw a line standing up," "Draw a line lying down," and "Draw what you think is number four," and "Draw a circle. Color the circle red." The present experimental version of the test contains ten tasks. Each item is scored for a number of specified credits totaling 22 points.

### *Results*

For the 121 pupils studied in this report, the mean score was 15.65 with a standard deviation of 4.71; the median was 16.75; and the modal score was 17. The reliability estimate based on use of the Kuder-Richardson formula 20 was found to be .86, and the standard error of measurement was 1.73. Each of four untrained raters was asked to score a random sample of 10 protocols. From these ratings, interrater reliability was estimated to be .99 through using the analysis of variance model (Kerlinger, 1971). Thus, the HEILRT test score was reliable in terms of the scorers' performance as well as the children's.

Validity was studied by considering the relationship of HEILRT scores to scores on the Metropolitan Readiness Test (MRT) (Hildreth, Griffiths, and McGauvran, 1965) given later in the kindergarten year and other variables. Table 1 shows the intercorrelations of these variables with the HEILRT. Since HEILRT correlated highly positively (.79) with MRT Total Score, it was a valid predictor of

TABLE I  
Correlations between HEILRT and Other Indices<sup>a,b</sup>

Variables	1	2	3	4	5	6	7	8	9	10	11	12
1. HEILRT												
2. Sex of Child (1= male, 2= female)	29											
3. Age	-02	-13										
4. Number of Siblings	-02	-04	-02									
5. Education of Mother	15	-05	01	01								
6. Education of Father	19	-06	03		47							
Metropolitan Readiness Test												
7. Word Meaning	50	09	19	-10	04	13						
8. Listening	52	05	19	-04	29	29	51					
9. Matching	65	07	01	-04	17	20	30	37				
10. Alphabet	66	23	-09	-13	14	30	38	40	40			
11. Numbers	63	16	08	09	23	31	52	53	42	60		
12. Copying	44	08	15	01	23	28	22	39	46	36	52	
13. Total MRT	79	16	11	09	25	35	64	71	68	74	86	69

<sup>a</sup> Decimal points omitted from correlation coefficients.

<sup>b</sup> Values exceeding .20 are significant at .05 level.



MRT total score. In addition, the test correlated positively with MRT subtest scores. It is important to note that neither MRT nor HEILRT correlated very highly with age, number of siblings, or educational level of the mother and father. Thus, the pattern of correlation would appear to lend some support to the construct validity of the measurement of the test by its independence from these variables and its high correlation with the MRT total score. However, the correlation of .29 between sex and HEILRT scores suggests that these scores were higher for girls than for boys. This outcome corresponds to the generally accepted notion that at younger ages, girls have higher readiness skills than do boys.

### *Discussion and Conclusions*

Although the HEILRT is still in experimental form, results suggest the test to be valid and reliable for readiness screening of young children. The ease of administration and scoring by untrained classroom teachers is an important advantage of the test. In addition, the test can be administered to groups of up to 30 children in only 15 to 20 minutes. It is thought that the test can be used most effectively as part of a schoolwide testing program for readiness when administered by the Guidance Department or reading teachers. It can be economically and easily given to all pre-school or kindergarten children in a school population. Those children who obtain low scores can be identified and examined further for specific readiness problems.

Another feature of the test is its obvious economy. The test takes little time to administer and to score. It does not require lengthy training for the examiner or scorer. Furthermore, the materials necessary are only a set of instructions, a set of crayons, and a piece of blank paper. Research on the test is continuing.

### REFERENCES

- Hildreth, G. H., Griffiths, N. L., and McGauvran, M. E. *Metropolitan Readiness Tests*. New York: Harcourt, Brace and World, 1965.  
Kerlinger, F. *Foundations of behavioral research* (2nd ed.). New York: Holt, Rinehart and Winston, 1973.

## INTERCORRELATIONS AMONG MEASURES OF INTELLIGENCE, ACHIEVEMENT, SELF-ESTEEM, AND ANXIETY IN TWO GROUPS OF ELEMENTARY SCHOOL PUPILS EXPOSED TO TWO DIFFERENT MODELS OF INSTRUCTION<sup>1</sup>

JOHN LEWIS

Winona State College

RICHARD ADANK

Winona Public Schools

The intercorrelations among intelligence, achievement, self-esteem, and anxiety measures were studied among fourth, fifth, and sixth grade pupils in self-contained and individualized programs. In addition to significant positive interrelationships among the measures of intelligence, achievement, and self-esteem for each of the groups exposed to different models of instruction, the lack of a significant negative correlation between the measure of anxiety and either the achievement or intelligence measure for the group exposed to individualized instruction as compared with the presence of a corresponding significant negative correlation for the group in the traditional self-contained model was judged noteworthy.

THIS study was undertaken to determine the intercorrelations among measures of intelligence, achievement, self-esteem, and anxiety among elementary pupils who were enrolled in two different styles of elementary instruction. The two schools were a traditional self-contained structure and an open school using the Westinghouse PLAN system of computer-backed individualized instruction.

### *Procedures*

The subjects were pupils in grade four, five, and six in both schools for whom complete data could be found. Intelligence was defined as

<sup>1</sup> The research efforts underlying this study were supported by Title III funds made available to the Winona Public School System, Winona, Minnesota, Project No. 33-71-4049 by the United States Department of Health, Education, and Welfare.

IQ scores from the SRA Tests of General Ability. Achievement was defined as the composite of raw scores from the Stanford Achievement Test, and self-esteem was measured by the Self-Esteem Inventory (Coopersmith, 1967). Anxiety was defined as the combined raw scores from the General Anxiety Scale for Children (Sarason, 1960) and the Test Anxiety Scale for Children (Sarason, 1960). These anxiety scales were constructed so that positive responses and hence higher scores indicate higher anxiety. The correlations were calculated among the variables at each of the three grade levels, then converted into Fisher's  $z$  scores, weighted by the  $N$ 's at each grade, summed, and then reconverted into  $r$  values.

### Results

The intercorrelations along with indications of significance for pupils in each of the two schools are shown in Table 1.

Within each group of pupils exposed to a different mode of instruction, significant positive intercorrelations were found among intelligence, achievement, and self-esteem measures and significant negative correlations were found between scores in the self-esteem and anxiety tests. Significant negative correlations appeared between scores in the anxiety measure and those in achievement and intelligence tests for pupils in the self-contained school, but non-significant negative correlations were found between scores in the anxiety

TABLE 1  
*Correlations among Intelligence, Achievement, Self-Esteem, and Anxiety Measures among Elementary Pupils in Two Models of Measured Instruction*

<i>Individualized Instructors Plan</i>				
<i>(N = 89)</i>				
Test Measure	1	2	3	4
1. SRA Tests of General Ability (IQ Scores)	—	.68**	.24*	-.16
2. Stanford Achievement Test (Composite)	.68**	—	.30**	-.16
3. Self-Esteem Inventory	.24*	.30**	—	-.23*
4. Combined Anxiety Score	-.16	-.16	-.23	—
<i>Self-Contained Structure</i>				
<i>(N = 130)</i>				
Test Measure	1	2	3	4
1. SRA Tests of General Ability (IQ Scores)	—	.59**	.34**	-.35**
2. Stanford Achievement Test (Composite)	.59**	—	.42**	-.38**
3. Self-Esteem Inventory	.34**	.42**	—	-.41**
4. Combined Anxiety Score	-.35**	-.38**	-.41**	—

\*  $p < .05$ .

\*\*  $p < .01$ .

measure and scores in the intelligence and achievement tests for pupils in the individualized school.

These results suggest that although the interrelationships among self-esteem, intelligence, and achievement measures did not vary substantially among pupils enrolled in the two models of instruction the use of the individualized model might reduce the extent of the negative relationship between anxiety and either achievement or aptitude.

#### REFERENCES

- Coopersmith, S. *The antecedents of self-esteem*. San Francisco: W. J. Freeman, 1967.
- Sarason, S. B. *Anxiety in elementary school children*. New York: Wiley, 1960.





## THE RELATIONSHIP BETWEEN ATTITUDES TOWARD SCHOOL AND ACHIEVEMENT FOR GROUPS OF ELEMENTARY SCHOOL CHILDREN EXPOSED TO TWO MODELS OF INSTRUCTION<sup>1</sup>

JOHN LEWIS

Winona State College

RICHARD ADANK

Winona Public Schools

A study of the relationship between attitudes toward school and learning among pupils in grades one through six in two different models of instruction revealed four significant correlations among twelve calculated. The implication was that attitudes towards school were not systematically related to learning among elementary pupils.

THIS study was undertaken to determine whether measured attitude towards school was related to an objective measure of learning within each of two groups of pupils exposed to different models of instruction. The first group attended an elementary school which afforded an open-styled model of instruction involving the Westinghouse PLAN computer-backed individualized program. The second group attended a traditional self-contained elementary school.

### *Procedures*

The subjects were 286 pupils in grades one through six in the individualized school and 335 pupils in grades one through six in the self-contained school. Both schools are located in Winona, Minnesota. Pupil attitude toward school was measured by a pictorial attitude

<sup>1</sup> The research efforts underlying this study were supported by Title III funds made available to the Winona Public School System, Winona, Minnesota, Project No. 33-71-4049 by the United States Department of Health, Education, and Welfare.

scale (Lewis, 1974) and pupil learning was measured by the Lee-Clarke Reading Test for pupils in grades one and two and by the Stanford Achievement Tests for pupils in grades three through six. The attitude measure was administered at the beginning of the school year. The achievement measures were given at the end of the school year.

### Results

The correlations between measured attitude and achievement at each grade level in each of the two schools emphasizing different models of instruction are presented in Table 1. It should be noted that the attitude scale provided for three possible responses: "like," "neutral," and "dislike." Very few pupils responded in the dislike category to any of the items. Thus, the scores on the attitude scale essentially reflected feelings of either liking or being indifferent toward school.

Significant relationships between attitude and achievement appeared at two of the six grade levels at each of the two elementary schools for a total of four significant correlations among the twelve groups. Non-significant relationships appeared among the other eight groups of pupils. Thus, responses of generally liking or being neutral toward the items on this scale taken at the beginning of the school year were not systematically related to pupil achievement in school measured at the end of the year.

### Discussion

The appearance of some significant relationships, however, did raise the question of why attitudes were related to achievement among certain groups of pupils but not among others. Since two of these significant correlations appeared among pupils in grades one and five in the school with an individualized program and among pupils in grades three and four in the school with a self-contained structure, it appears

TABLE 1  
*Correlations between Attitudes toward School and Learning  
in Two Models of Elementary Instruction*

Model/Grade	Grade Level					
	1	2	3	4	5	6
Individualized	.26*	-.13	.11	.00	.35*	.20
Self-Contained	-.06	-.12	.40**	.30*	.15	.22

\* Significant at the .05 level

\*\* Significant at the .01 level.

that factors contributing to the question of a relationship between these two variables would *not* seem to include either grade level or model of instruction. It was previously pointed out that subjects generally responded in terms of a neutral or positive attitude toward school. Thus, these findings should not be generalized to situations in which a large number of pupils exhibit negative feelings towards school.

#### REFERENCE

- Lewis, J. A pictorial attitude scale for elementary pupils. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1974, 34, 461-462.



## CONCURRENT VALIDITY OF THE TORRANCE TESTS OF CREATIVE THINKING AND THE WELSH FIGURAL PREFERENCE TEST

THOMAS M. GOOLSBY, JR AND LOREN D. HELWIG  
University of Georgia

For a sample of 79 fifth grade pupils from a semi-rural farming area in the Southeastern United States intercorrelations among subtest scores and total scores of the Torrance Tests of Creative Thinking (TTCT) and the Welsh Figural Preference Test (WFPT) revealed little if any relationship either between subtests from the TTCT and those from the WFPT or between the two total tests themselves. It would appear that the two scales which were designed to measure constructs related to creativity share little common variance.

THE Torrance Tests of Creative Thinking (TTCT) and the Welsh Figural Preference Test (WFPT) are two tests purporting to measure constructs related to creativity (Torrance, 1974; Welsh, 1959, 1972). Whereas the TTCT scales appear to be more cognitively oriented in approach, the WFPT seems to reflect affective characteristics.

The purpose of this study was to explore the patterns of interrelationships among the part scores and total scores of the TTCT and WFPT for a sample of fifth grade pupils.

Seventy-nine fifth grade pupils from a semi-rural farming area in the Southeastern United States were administered the Figural Form B of TTCT and the WFPT within two days.

The Figural Form B of the TTCT yields subscores of Fluency (*Fl*), Flexibility (*Fx*), Originality (*Or*), and Elaboration (*El*). The WFPT furnishes subscores on subtests named "origence" (*Og*) and "intelligence" (*In*).

A zero-order intercorrelation matrix with means and standard deviations was obtained using the total test and subtest scores for the TTCT and WFPT.



TABLE I  
*Intercorrelations among Subtest and Total Test Scores with Means and Standard Deviations for TTCT and WFPT (N = 79)*

	<i>Fl</i>	<i>Fx</i>	<i>Or</i>	<i>El</i>	<i>In</i>	<i>Og</i>	<i>M</i>	<i>SD</i>
TTCT (total test)					-15	07	144.16	41.61
Fluency ( <i>Fl</i> )							18.48	6.00
Flexibility ( <i>Fx</i> )	76						13.66	4.00
Originality ( <i>Or</i> )	14	17					29.56	13.50
Elaboration ( <i>El</i> )	49	42	40				82.28	28.90
WFPT (total test)	11	-05	-02	02			75.43	18.70
"Intellectence" ( <i>In</i> )	-08	-12	-05	-17			22.03	4.90
"Origence" ( <i>Og</i> )	14	-02	-01	07	20		53.41	17.10

Note.—Decimals omitted for correlation coefficients.

Table I shows the intercorrelations among subtest and total test scores with means and standard deviations for TTCT and WFPT. The correlation between total test scores for TTCT and WFPT was .02. Evidence of concurrent validity between TTCT and WFPT was not demonstrated for fifth grade pupils. That this evidence was disconfirming was not a surprising outcome, since the TTCT and WFPT were constructed from different frameworks of reference and/or theory bases.

Some correlative evidence was viewed as useful in identifying sources of shared variance in scores among selected pairings of subtests or of subtests with an entire test of which they were not parts.

## REFERENCES

- Torrance, E. P. *Torrance Tests of Creative Thinking*. Lexington, Mass.: Personnel Press, 1974.
- Welsh, G. S. *Welsh Figural Preference Test*. Palo Alto, Calif.: Consulting Psychologists Press, 1959.
- Welsh, G. S. A two-dimensional personality model for research in social science. *Research Previews*, Institute for Research in Social Science, The University of North Carolina, Chapel Hill, Vol. 19, No. 1, April, 1972.

## THE CORRELATION OF PARTIAL AND TOTAL SCORES OF THE SCHOLASTIC APTITUDE TEST OF THE COLLEGE ENTRANCE EXAMINATION BOARD WITH GRADES IN FRESHMAN CHEMISTRY

L. G. PEDERSEN

University of North Carolina, Chapel Hill

For a sample of 325 students enrolled in freshman chemistry, it was found that performance on the Verbal and Mathematics Sections of the Scholastic Aptitude Test of the College Entrance Examination Board exhibited low predictive validity in relation to a criterion of grades in freshmen chemistry.

A major problem at many universities is how to effect curriculum changes which will increase the chances of college success for educationally deprived and underprepared students (Kotnik, 1974; Meckstroth, 1974; Wartell, 1974). The Chemistry Department at the University of North Carolina, Chapel Hill, responded to this need by scheduling a preparatory course for freshman chemistry for Fall, 1974. An immediate concern pertained to the criterion by which a student would be placed in such a preparatory course. Would one use as a predictor, grades on the Scholastic Aptitude Test (SAT) of the College Entrance Examination Board (CEEB), high school rank, a linear combination of these two variables, or some other procedure?

### *Purpose*

The purpose of this study was to determine the degree of relationship between level of achievement in freshman chemistry and (a) scores on the Verbal (*V*) part of the CEEB, (b) scores on the Mathematics (*M*) part of the CEEB, and (c) an unweighted composite (*T*) of the two part scores. In addition the worthiness of a model

involving use of cutoff scores was evaluated for its predictive capability.

### *Method and Findings*

In the present study a data base was established in terms of the grades in the first semester freshman chemistry (4.0 scale; A = 4.0, B = 3.0, C = 2.0, D = 1.0, and F = 0.0) of 325 students enrolled during Fall 1974 and the Mathematics (*M*), Verbal (*V*), and Combined (*T*) scores of the same students. The data base consisted of two large sections taught by two different teachers. About 15% of the students had previously been placed in an "honors" type section on the basis of superior preparation, high school standing, and special examinations. To construct a criterion for prediction, the average SAT scores for student grades in the five grade categories (Table 1) were evaluated. For all three SAT categories—*V*, *M*, and *T*—the *average* grade point was seen to increase regularly as the scores in the three test categories increased. If one wishes to use the SAT score averages to predict grades, one must establish a cutoff for A's, B's, or other grades in the three SAT categories. A logical way to decide between an A and B would be to choose as a cutoff the midpoint between the A and B score—those above this midpoint would receive A's, those below (but above the C/B midpoint) would receive B's. These cutoffs were computed for the three SAT categories and are given in Table 2.

Also given in Table 2 are the results for the predictability of the model. The average cutoffs suggested in Table 2 were used to recalculate the grades for the 325 students of the data base. The level of predictability is seen to be almost useless for forecasting the actual grade. The model works considerably better for predicting the grades within one letter grade of the correct grade (last line of Table 2), but it may not be very satisfying to a student or advisor to know that the stu-

TABLE 1  
*Mean and Standard Deviation (SD) of Verbal (V), Mathematical (M), and Total (T) Scores of the Scholastic Aptitude Test of The College Entrance Examination Board for Freshman Students in Chemistry Receiving Different Grades*

Grade	<i>V</i>		<i>M</i>		<i>T</i>	
	Mean	SD	Mean	SD	Mean	SD
A	549.1	83.1	622.1	62.8	1171.2	128.5
B	526.8	70.7	577.0	59.2	1103.8	114.0
C	501.1	76.2	559.5	58.6	1060.7	110.8
D	479.8	60.6	534.8	60.9	1014.6	90.5
F	467.7	54.8	514.8	68.3	982.5	105.5
All Students	510.3		566.5		1076.8	

Note—The average grade point in Fall, 1974 for all 325 students in the data base was 2.29

TABLE 2  
Cut-off Grading Model Based on SAT Scores

Decision Between:	<i>V</i>	SAT Category <i>M</i>	<i>T</i>
A and B	537.9	599.6	1137.5
B and C	513.9	568.3	1082.3
C and D	490.5	547.2	1037.7
D and F	473.7	524.8	998.6
Percentage of Accurate Grade Predictions <sup>a</sup>	24.0	28.6	31.4
Percentage of Nearly Accurate Grade Predictions <sup>b</sup>	56.2	64.7	66.2

<sup>a</sup> Percentage of grades predicted correctly on basis of cut-off model. Thus only 31.4% of the 325 students had their grades predicted correctly on basis of composite (*T*).

<sup>b</sup> Percentage of grades predicted correctly within one letter grade, i.e., F-D, D-C-B, C-B-A, B-A on basis of cut-off model. Thus 66.2% of the 325 had their grades predicted within one letter grade.

dent is predicted to make a D, C, or B with 2/3 probability. In all cases the combined *T* category was slightly superior to the *M* category, and the *M* category was superior to the *V* category. The correlation coefficient was calculated for each SAT category vs. grade distribution. The coefficients for *V*, *M*, and *T* measures were, respectively, .32, .42, and .43, all of which were significant beyond the .01 level. The magnitudes of the correlation coefficients indicated that the level of correlation was not very useful just as the grade cut-off model indicated.

### Conclusions

The following conclusions were drawn:

1. The functional dependence of average grade obtained in freshman chemistry vs. average SAT category was regular and monotonic, i.e., the higher the average scores on the *V*, *M*, or *T* measures for a group of freshmen, the higher would be the average grade of that group in freshman chemistry. The standard deviations, however, were large.
2. There was little freshman chemistry grade predictability from standing in any of the SAT-grade categories.
3. Use of SAT scores for deciding the placement of an individual student was not statistically justified.

### REFERENCES

- Kotnik, J. What is being done to help the underprepared student? *Journal of Chemical Education*, 1974, 51, 165-167.
- Meckstroth, W. H. A chemistry course for underprepared students. *Journal of Chemical Education*, 1974, 51, 329.
- Wartell, M. A. Chemistry and the educationally disadvantaged student. *Journal of Chemical Education*, 1974, 51, 116.





## BOOK REVIEWS

Jere E. Brophy and Thomas L. Good. *Teacher-Student Relationships: Causes and Consequences*. New York: Holt, Rinehart & Winston, Inc., 1974. Pp. xvi + 400. \$5.95 (paperback).

This book is a well-written review and discussion of research on teacher expectations and attitudes and how they affect and are affected by various characteristics of students. The approximately 500 references cited in the 11 chapters reveal something of the volume of research stemming from Rosenthal and Jacobson's *Pygmalion in the Classroom*. As is frequently the case in educational research, the results of these efforts are open to diverse interpretations. It is obvious from reading the book that its authors are committed to the belief that teacher expectations do, under certain circumstances and with certain teachers and students, affect the behavior of the latter.

Since, at least for this book, a listing of the table of contents reveals the substance of the topics considered, here it is: 1. Individual Differences in Teacher-Student Interaction Patterns; 2. Teacher Expectations; 3. Studies of Experimentally Induced Expectations; 4. Naturalistic Studies of Teacher Expectation Effects; 5. The Influences of Teachers' Attitudes toward Students on Classroom Behavior; 6. Teacher Interview and Questionnaire Studies; 7. The Influences of the Sex of the Teacher and Student on Classroom Behavior; 8. Individual Differences and Their Implications for Teachers and Students; 9. Promoting Proactive Teaching; 10. Classroom Research: Some Suggestions for the Future; 11. Implications for Teaching.

As concluded from the summaries in Chapters 3 and 4, more of the studies of experimentally induced expectations found negative results, but the reverse is true of naturalistic studies. Although the authors are aware of the comparative advantages and disadvantages of experimentation and naturalistic (observation, correlation) studies, they show a definite preference for naturalistic studies. In fact, citing problems of adequate controls and ethics, they reiterate throughout the book their feeling that no more experimental studies should be conducted to demonstrate the reality of expectation effects. Another point which they continually stress is that most teaching is "reactive" rather than "proactive."

Since this book is really about affective influences in education, attention is not restricted to teacher expectations. It is recognized that expectations are not independent of attitudes, and that the more

general question is how teacher and student attitudes and behavior interact to influence each other. In discussing the results of variables related to this interaction, some interesting conclusions are drawn, for example:

1. Teachers prefer conforming, orderly students to assertive, independent students.
2. Teachers form strong first impressions of students on the basis of their early contacts with them.
3. Teachers are usually not very aware of their own behavior toward students or how it affects the latter.
4. There is little or no evidence that male teachers favor male students or that female teachers favor female students in any general sense.
5. Male teachers appear to be more achievement-oriented than female teachers, and hence more interested in putting across the material than in determining whether students understand it.
6. The more similar a teacher and student are, the more likely the teacher is to like the student.

The authors do not stop at reviewing and discussing the research literature. In Chapter 9 they give a number of concrete prescriptions for promoting "proactive teaching," that is, making teachers aware of their behavior in the classroom and helping them change it. Furthermore, Chapters 9 and 10 contain many interesting suggestions for research on expectations and attitudes. Finally, the implications for teaching listed in Chapter 11, especially those concerning the differential treatment of students toward whom the teacher has low or high expectations, are useful even if they admittedly go beyond the research data.

Basically, what the authors suggest in the last few chapters is that teaching should be appropriate to the needs and personality of the learner. Of course, in a class of 30 students, and at the secondary school level several classes of 30 students, this is easier said than done. The usual situation is that some teachers manage to meet the needs of some students—more frequently the bright, conforming, attentive, white, middle-class students who are seated in the front or middle row. As the book makes clear, teachers trying to put across the lesson are not very sensitive detectors of what students know or are thinking. And, as anyone who has tried it knows, sensitivity training and behavioral feedback are not invariably effective, even when conducted by an expert with highly motivated teachers.

Perhaps the statement in the preface of the book that "Despite years of educational research, relatively little is known about the characteristics of effective teachers or the behaviors involved in effective teaching," is too pessimistic in the light of the material covered in this book. To be sure, research on the characteristics and behavior of effective teachers should continue. But the problem of effective

teaching appears to be less one of lack of knowledge of what to do and more a matter of how to do it. Brophy and Good have written a book that includes a good summary of research on teacher expectations and attitudes and variables associated with them. The book, however, is not limited to a description of the what of effective teaching; it also makes cogent suggestions as to how it should be done. As such, the book deserves serious attention by teachers, student-teachers, and educational researchers. Although not a textbook in the strict sense, it can be used with profit as associated reading in courses on teaching methods, educational psychology, and curriculum.

LEWIS R. AIKEN, JR.

Oscar K. Buros (Ed.). *Tests in Print II*. Highland Park, N. J.: Gryphon Press, 1974. Pp. xxxix + 1107. \$70.00. (no postage charges on prepaid orders.)

This extremely useful volume is dedicated to the consistent reviewers of tests where consistent means contributing reviews to each of the seven *Mental Measurement Yearbooks*: Anne Anastasi, Howard R. Anderson, Walter V. Kaulfers, Victor H. Noll, and Arthur E. Traxler. Following Oscar Buros' summary closely, the book lists 2,467 tests in print as of early 1974; 16,574 references through 1971 on specific tests; a directory of 493 test publishers with complete listing of their tests; a specific author index for each test with references; a title index which includes both in-print and out-of-print tests; a comprehensive cumulative author index to approximately 70,000 documents (tests, reviews, excerpts, and references) in *Tests in Print II*, the seven *Mental Measurements Yearbooks*, *Personality Tests and Reviews*, *Reading Tests and Reviews*, and a scanning index for quickly locating test designed for a particular population. Also included in the volume is a reprinting of the 1974 APA-AERA-NCME *Standards for Educational and Psychological Tests*.

The "Expanded Table of Contents" presents a complete list of all categories under which tests have been classified. The number cited are test numbers, not page numbers. For example 323 pertains to the first in the list of group intelligence tests while 483 pertains to the first listed individual intelligence tests. The number 323 to 482 appear in brackets above the title of each of the group intelligence tests. Such tests are listed in alphabetical order by title. Above the title *College Board Scholastic Aptitude Test* appears [357] immediately followed by a paragraph descriptive of its function, testing candidates for college entrance, its dates 1926-73, the acronym SAT, its administration on specified dates established by the publisher, ETS, and its two scores verbal and mathematical and the citations of reviews and references in *Mental Measurement Yearbooks* 4, 5, 6, and 7. This is followed by 148 complete and accurate references numbered from 420 to 567 in the

*order of their publication.* This is succeeded by a "cumulative name index" with the names of authors in *alphabetical order*. This enables a reader to locate quickly a reference from knowledge of the name, or names of its authors. (The same name may pertain to more than one reference, e.g., Brigham C. C.: 420-3, 427, 432.)

The PREFACE of *Tests in Print II* gives a brief history of *Tests in Print I* and along with the chapter headed INTRODUCTION presents a history of the seven *Mental Measurements Yearbooks* from the *Nineteen Thirty Eight* and *Nineteen Forty Mental Measurements Yearbooks* to the *Seventh Mental Measurements Yearbook* and MMY monographs *Reading Tests and Reviews* (RTR) and *Personality Tests and Reviews*. Seven additional monographs are planned for 1975: English, foreign languages, intelligence, mathematics, science, social studies, and vocations. One cannot but be awed by the volume of business contemplated by Oscar Buros, his small staff, and his numerous contributors.

In the INTRODUCTION are seven interesting and informative tables: Table 1 shows the numbers and percentages of tests listed in TIP I (1961) and in TIP II (1974) as classified by the EXPANDED TABLE OF CONTENTS inside the front cover and rear covers of TIP II. The total numbers of tests are 2,126 (1961) and 2,467 (1974). Table 2 reports the numbers and percentages of the 2,467 tests thus classified which are new or revised. Table 3 gives the numbers and percentages of tests classified by countries. In 1974, the United States had 2,204 or 85.3%, Great Britain 181 or 7%. South Africa, Australia, and Canada has 64, 53, and 50 tests 2.5%, 2.1%, and 1.9% respectively. Table 4 gives the numbers and percents of reviews, excerpts, and references as categorized by the EXPANDED TABLE OF CONTENTS. Table 5 lists the titles of the tests with 100 or more references through 1971. The *Rorschach*, the *Minnesota Multiphasic* (MMPI), the *Thematic Apperception Test* and the *Stanford-Binet Intelligence Scale* lead with 4,578, 3,855, 1,765, 1,408 references respectively. These are closely followed by the *Edwards Personal Preference Schedule*, the *Strong Vocational Interest Blank for Men*, and the three *Wechsler Intelligence Scales*. Table 6 gives the titles of tests with 23 or more references in the years 1969-71. As noted above TIP II includes the APA-AERA-NCME *Standards for Educational and Psychological Tests* as Published in 1974. TIP II concludes with a PUBLISHERS DIRECTORY AND INDEX, an INDEX OF TITLES, an INDEX OF NAMES, and a SCANNING INDEX. "THE SCANNING INDEX will probably be most useful in helping readers locate all tests in a particular area which are suitable for a given population . . ." (p. xxxv). The introductory chapter concludes with advice on how to use TIP II and specification of the Information presented, when available, about each test. The chapter concludes with a statement of the objectives of Oscar Krisen Buros and his colleagues in The Institute of Mental



Measurements. As the president of Educational Testing Service, William Turnbull, said about him "If Oscar Buros didn't exist, we would have to create him."

MAX D. ENGELHART

Jeremy D. Finn. *A General Model for Multivariate Analysis*. New York: Holt, Rinehart and Winston, 1974. Pp. xviii + 589. \$10.95.

Some books remind one of a flock of sheep in a spring meadow: constantly leaping off in 20 directions but always, amazingly, a unified whole no part of which gets lost. Finn has written just such a book. He is concerned with multivariate analysis as a tool for conceptualizing and understanding behavioral phenomena. He is equally concerned with fitting algebraic models to multiple random (outcome) variables. He explores such models in multivariate multiple correlation and regression, canonical correlation, principal components, multivariate analysis of variance and covariance, and discriminant function analysis. He is also involved with formal aspects of estimation and hypothesis testing, including topics as arcane (from the point of view of behavioral scientists) as estimability. He strives to represent all data, all transformations, and all computations in standard matrix notation. He uses five large-scale, real-life, multivariate (largely behavioral) problems to illustrate relevant research areas, exemplify computations, and tell the research story from initial conceptualization and design to final inference and interpretation. He discusses computer techniques and devotes a lengthy appendix to the crucial process that turns stacks of output into readable and accurate results and discussion in a journal article. Almost as a sideline, Finn teaches more matrix algebra than is contained in many mid-level texts. Finally, he provides timeliness by devoting considerable space to topics such as reparameterization of analysis of variance models which until now have been poorly or incompletely treated in the literature. Surprisingly, this unbelievable potpourri works very well.

The secret of Finn's success is the discipline he brings to his presentation. He provides the reader with an overview of the techniques to be covered, he discusses in detail the sample problems which will be used to exemplify them, and then he sets about providing solutions for the problems one step at a time.

Finn's first sample problem (based upon data collected by I. Leon Smith) is a small classic. The topic is the relationship between creativity and divergent achievement. Divergent achievement is defined as a quantitative indicator of synthesis and of evaluation as determined by the tests of Kropp, Stoker, and Bashaw. The major question asked is, "... whether an individual's level of creativity is a determinant of divergent achievement, and further, whether this contribution represents an effect that cannot more parsimoniously be at-



tributed to general intelligence" (p. 11). Creativity levels are measured through three of Guilford's subtests: *consequences obvious*, *consequences remote*, and *possible jobs*. Intelligence is measured on the Lorge-Thorndike Multi-Level Intelligence Test (G-1). The two hypotheses to be tested are: (1) that the three tests of creativity and the two of divergent achievement represent a homogeneous set of cognitive processes; and (2) that levels of creativity determine to a considerable degree an individual's divergent achievement functioning. The first hypothesis is investigated through principal components analysis, and the second via multivariate multiple regression. As an added stroke of sophistication, Torrence's postulate that creativity has a greater effect when coupled with higher intelligence is operationalized through three additional independent variables: the interactions of the intelligence variable with the measures of divergent achievement. The inferential part of the multi-variate analysis of variance is conducted through logically ordered, sequential tests, and the latter three predictors are therefore relegated to the last positions in the predictor vector.

The other four sample problems are a word-memory experiment, a study of dental calculus reduction, an essay grading study, and an experiment on programmed instruction effects.

The first step in solving the sample problems is to set them up in matrix form. To demonstrate this process Finn begins with matrix notation, progresses through simple operations such as multiplication and on to more complex ones including orthonormalization and Kronecker products, and concludes with the Cholesky factoring procedure and inversion techniques. Related procedures for calculating determinants, finding characteristic roots and vectors, and taking derivatives of linear functions are described. There are two faults with Finn's presentation of this material. First, he continues to use vector representation, presumably for the sake of clarity, far past the point where matrices would represent not only a more satisfactory notational condensation but also a much more satisfying conceptual framework. Secondly, several of the computational procedures are outlined in a way that keeps them close to the related computer routine, but robs them of the inherent simplicity that other algorithms would provide. The Cholesky factoring procedure is a case in point.

Before proceeding to multivariate multiple regression analysis, Finn pauses to discuss multivariate data summarization procedures. He introduces the expectation and variance operators, variance-covariance matrices, standardization, the multivariate normal distribution, and conditional distributions and expectations. Considering the compactness of this section, the treatment is good, although once again it would probably have benefitted from more matrix representation and fewer vectors. In this chapter, as he does throughout the book, Finn sets off illustrative computations in half-tone boxes. Many of these ex-

amples are excellent, but some are so compressed that they lose their effectiveness.

One very worthwhile example demonstrates clearly the way in which failure to take into account mean differences among subgroups can profoundly bias correlations among variables.

Chapter 4 deals with multivariate multiple regression analysis. The framework of this chapter serves as a blueprint for the treatment of the other techniques covered in the book. The univariate model is set forth, expressed in matrix terms, and a simple example provided. Next the estimation of parameters is discussed and conditions for estimability listed. Dispersion and prediction in the univariate case follow. The multivariate model is developed in a like sequence. Computational forms are set forth, and the first sample problem is employed to illustrate the technique in detail. Tests of significance are presented in the following chapter and exemplified again with the creativity-achievement sample problem. The significance tests include not only univariate statistics but also the likelihood ratio test and various approximations to it.

Step-down analysis is virtually the only approach to model-building offered, and, while it provides a satisfying rigorous test procedure and avoids the quicksand of various error-rates, it is a definite departure from Finn's otherwise down-to-earth presentation of practical procedures. The sample problem allows for a logical ordering of independent variables, but many research problems do not, and step-down analysis (basically sequential orthogonalization) is virtually useless when a priori logical ordering cannot be specified.

For those of you who are curious, analysis of the first sample problem yields beautifully straightforward results. In this sample of observations, only general intelligence plays a significant role in divergent achievement. There does not seem to be a differential creativity effect either across intelligence levels or at any particular level. The univariate results show that "... synthesis is more affected by intelligence than is evaluation. The stepdown statistics indicate that in fact the more complex trait, evaluation, does not contribute to the association with the predictors. The relationship between the two sets of measures is parsimoniously summarized in the correlation (+.64) of the two simplest constructs, intelligence and synthesis" (p. 17).

Chapter 6 deals with correlation: simple, partial, multiple and canonical. A number of useful tests are discussed and several little-known but worthwhile procedures, such as the Olkin and Siotani test for the correlation of each of two variables with a third, are detailed. The very considerable difficulties involved in the interpretation of canonical correlations are well illustrated. The section on use of principal components, however, is much too compressed both conceptually and mathematically. Moreover, many factor analysts might argue with a substantial part of Finn's interpretation.

The material in chapters 7 through 10 on multivariate analysis of variance and covariance is excellent. There are actually too many good features to attempt to list them all. Some of the more interesting and useful ones deserve at least brief mention, however. The material on the use of Kronecker's delta to generate comparisons has probably never been better presented. The treatment of reparameterization is very thorough. It includes selection of contrasts, construction of contrast matrices, bases, estimability of interactions, and computations. The idea of nesting is expressed simply but correctly. The sample problems are apropos and generally well handled.

The segment in chapter 10 on discriminant function analysis is abbreviated in comparison to the other topics Finn covers, and appears to have been added almost as an afterthought. The dental calculus reduction study used as an example problem is a less than ideal choice: it lacks clarity and emphasis.

One or two additional points should be mentioned. The book contains no exercises except for a set to accompany the chapter on matrix operations. The bibliography is not bad, but is neither comprehensive nor especially representative of the major literature in applied multivariate analysis. The very lengthy appendix on the MULTIVARIANCE program is useful as a guide to interpreting and using complex computer output, but it is unduly bulky: it could have been reduced by two-thirds with little sacrifice in utility. As it stands, it gives more than a bit of the impression of being a long advertisement for a particular canned program.

In summary, one must conclude that Finn has written an unusual but uniquely useful book. It does not, of course, provide anything as ambitious as a formal general model for multivariate statistical analysis. What it does do is weave the many strands of applied multivariate procedures including problem conceptualization, formal models, matrix representation and computations, and statistical estimation and hypothesis testing together into a single fabric that connects real-life multivariate behavioral research with the best in analytical procedures. *A General Model for Multivariate Analysis* will be of use to many behavioral scientists and should achieve considerably popularity among them.

JAMES A. WALSH  
*University of Montana*

Gerald L. Isaacs, David E. Christ, Melvin R. Novick and Paul H. Jackson. *Tables for Bayesian Statisticians*. Iowa City, Iowa: The University of Iowa, 1974. Pp. 377. \$15.00 (paperback).

This new collection of tables was prepared at the Lindquist Center for Measurement, The University of Iowa. Its appeal to Bayesian statisticians arises in the choice of distributions and in the method of

tabulation. There are 20 tables devoted to eight distributions, together with another 24 tables of related functions and transformations.

Two introductory sections describe the usage of the tables, the notation, some definitions, some identities, some remarks on interpolation, and some examples. Further references are given to the recent text of Novick and Jackson (1974).

The first tables include the usual reciprocals, squares, cubes, square roots and cube roots, positive and negative exponentials, and common and natural logarithms. Next are a table of natural logarithms of  $n$ -factorial and a table of binomial coefficients. These are followed by tables for converting degrees to radians and radians to degrees. This is done separately for degrees with decimals and degrees with minutes. The trigonometric functions consist of the sine, the tangent, and the square of the sine function, together with tabulations of the inverses of these three functions.

The next two tables are the hyperbolic tangent, and its inverse: the hyperbolic arctangent. The latter is the Fisher  $z$ -transformation. This transformation applied to a correlation coefficient produces a variable which may be almost normally distributed.

A table of the log-odds transformation is given, that is

$$\delta = Ln[\pi/(1 - \pi)]$$

is given, together with a table of the inverse transformation. If the prior distribution for parameter  $\pi$  is assumed to be Beta, then the posterior distribution for  $\delta$  is nearly normal.

The cumulative normal distribution is tabulated, together with a table of percentage points (critical points), and with 10 pages of random observations (random normal deviates). The Student  $t$  distribution is tabulated in full for degrees of freedom one through 25, 40, and 60. A separate table is given for the percentage points. Sets of 500 random observations drawn from the Student  $t$  distribution are given for each of 18 choices of the number of degrees of freedom.

For the Chi-Square distribution the tabulation is by percentage points for various degrees of freedom. The 25 percentage points chosen cover both tails of the distribution as well as the center. A table is given of the end points of the interval of highest density for various specified probabilities and for specified degrees of freedom. In addition there are sets of random observations. The Inverse Chi-Square distribution and the Inverse Chi distribution are listed in three tables each, following the format described for the Chi-Square distribution.

The  $F$  distribution is tabulated by percentage points. The degrees of freedom for both numerator and denominator range up to 100, while the percentage points are chosen to give coverage in the center as well as in the tails. Next, a short table of Behrens-Fisher percentage points is given. Finally, there are three lengthy tables devoted to the Beta distribution. The first gives percentage points, the second gives intervals



of highest density, and the last gives probabilities that one Beta variable is greater than a second.

A concluding section presents the methods used to construct the tables, together with references to the computer subroutines used.

Within the tables the values of the arguments are closely spaced. In most cases a comfortable number of significant digits are provided. Though differing in scope from the classic Biometrika Tables and the statistical tables and Fisher and Yates, the reviewer feels that the current volume may well approach them in importance. It should be noted that the funding for this extensive project was through the generosity of the Iowa Testing Programs of the University of Iowa.

### REFERENCE

Novick, M. R. and Jackson, P. H. *Statistical methods for educational and psychological research*, New York: McGraw-Hill, 1974.

THOMAS CHURCH  
*Governors State University*  
*Park Forest, South Illinois*

David A. Payne. *The Assessment of Learning: Cognitive and Affective*. Lexington, Mass.: D. C. Heath and Company, 1974. Pp. 524. \$9.95.

In the words of its author, this book was written "to provide the ever-increasing number of classroom teachers and professional evaluators a practical and efficient set of techniques to aid in evaluating learning outcomes." From the preface it is assumed that this aim will be realized primarily by catching these two groups of individuals when they are undergraduate or graduate students in courses on "test construction or evaluative research methods."

The author of this book is Professor of Educational Psychology at the University of Georgia and Review Editor of the *Journal of Educational Measurement*. The book, a revision and expansion of his earlier paperback volume (Payne, 1968), is chock-full of facts and methods pertaining to educational assessment. Within its 524 pages are 18 chapters, several appendices including some useful statistical tables, a glossary of terms, and the usual prefatory and index material. The chapters range in length from 12 to 46 pages, and are divided into six sections: I. An Overview (Chapt. 1), II. Planning for Instrument Development (Chapts. 2-4), III. Instrument Development (Chapts. 5-8), IV. Summarizing and Interpreting Test Performances (Chapts. 9-10), V. Instrument Refinement (Chapts. 11-12), VI. Other Sources and Uses of Assessment Data (Chapts. 13-18).

Cognitive objectives of learning are emphasized in the book, but somewhat unusual and timely is the comprehensive treatment of affective and psychomotor objectives. For example, five separate chapters



are devoted to the specification and measurement of the affective outcomes of education. The book possesses several other unusual features, but whether or not they are meritorious depends on one's point of view. It could be argued that the design and style of the chapters—summaries at the beginning rather than the end; many lists, tables, flowcharts, etc.; technically competent but cumbersome and unflowing prose—is too reminiscent of journal articles in APA format. Personally, I found the emphasis on structure and handbook-like detail somewhat unstimulating and distracting at times. Whether fledgling students who frequently go through statistically-oriented courses in a hazy-dazy condition will appreciate this style is guesswork. Students do not invariably prefer the same textbooks as their professors, but it is usually wise to err on the simple side!

Even with its shortcomings, there are numerous positive features to this book. Chapters 2 and 3 present a thorough coverage of educational objectives, Chapter 5 contains many helpful illustrations of good and poor item writing, and Chapter 8 gives excellent descriptions of self-report affective items and inventories. Chapter 10 on test interpretation, Chapter 13 on criterion-referenced measures, and Chapter 16 on the assessment of affective, performance, and product outcomes by direct observation are also noteworthy. I was also delighted to see a glossary in the appendix.

On the other hand, rather than taking up so much space in Chapter 14 with critical reviews of standardized achievement tests, it would probably have been more helpful to harassed teachers if the author had simply given his recommendations as to which tests are most appropriate for particular situations and some comments on the uses of standardized achievement tests in the schools. Furthermore, I cannot imagine that prospective teachers would be interested in all of the details of specific instruments presented in Chapter 15.

Among the topics receiving little or no attention, but deserving more, are the use of tests in accountability and performance contracting, gain scores, ethical and ethnical issues, the use of tests in prescriptive teaching as well as diagnosis of learning difficulties, and formative vs. summative evaluation. Chapter 9 is a good overview of statistics for testing, and unlike some authors this one did not confuse the definitions of percentile and percentile rank. Unfortunately, he was inconsistent in his definitions of variance and standard deviation: in the summary it's  $N$ , and in the chapter it's  $N-1$  in the denominator. A set of problems and exercises, especially in the more statistical chapters, would also have been nice.

Basically, David Payne has written a combination "how to do it" and reference book on assessment and evaluation for educators. He has done a good job of representing the current state of the art in educational testing. The book will serve well as a handbook or resource book for teachers and professional evaluators, and should

also have its share of adoptions for courses on "educational," if not "psychological," measurement and testing. In designing a textbook for a first course in educational testing, however, the author and editor would have fared better if they had given more thought to motivational features. As the author would undoubtedly agree, affect, as well as cognition, must be taken into account by educators of every stripe—including textbook writers.

### REFERENCE

Payne, D. A. *The specification and measurement of learning outcomes*. Lexington, Mass.: Blaisdell, 1968.

LEWIS R. AIKEN, JR.

David A. Payne and Robert F. McMorris (Eds.). *Educational and Psychological Measurement, Contributions to Theory and Practice*. (2nd. ed.) Morristown, N. J.: General Learning Corporation, 1975. Pp. xx + 397. \$6.50 (paperback).

The first edition of this outstanding anthology provided students and teachers "convenient and readily accessible summaries of concern to the undergraduate or graduate student of educational and psychological measurement." Comparison of both editors reveals that the editors have made numerous changes from the first edition to the second. Twenty-eight of the quoted selections are new while eighteen have been retained from the first edition. A few of the selections new to this anthology, actually appeared years ago. These include the late Percival Symond's "Factors Influencing Test Reliability (1928)." Stephen Corey's "Measuring Attitudes in the Classroom (1943)" and Anne Anastasi's "Some Implications of Cultural Factors for Test Construction (1949)."

Among the major contributions appearing in both editors are "Measurement and the Teacher," "Evaluating Content Validity," and "The Social Consequences of Educational Testing" by Robert L. Ebel; "Response Sets and Test Design" by Lee J. Cronbach; "Convergent and Discriminant Validity" by Donald T. Campbell and Donald W. Fiske; "Guidelines for Testing Minority Group Children," by Martin Deutsch and others, and this reviewer's "Suggestions for Writing Achievement Test Exercises," and "Non-Apparent Limitations of Normative Data" by Junius A. Davis. Among the influential selections appearing in the second edition only are "The Validation of Educational Measures," and "Course Improvement Through Evaluation" by Lee J. Cronbach, "Concepts of Achievement and Proficiency," by William E. Coffman and "Implications of Criterion-Referenced Measurement," by W. James Popham and T. R. Husek. Also influential and timely are "Testing for Accountability" by Ralph

W. Tyler, "Testing Hazards in Performance Contracting" by Robert E. Stake; "Expectancy Tables—a Way of Interpreting Test Validity" by Alexander G. Wesman.

Other commendable features of the second edition of this anthology are a list of 24 references "Current Textbooks in Tests, Measurement, and Evaluation," a "Textbook Reference Chart" which correlate the chapters or selected articles in the second edition of this anthology with the relevant chapters of the 24 textbooks. The anthology concludes with a bibliography of 334 references and an excellent glossary of 148 measurement terms.

In concluding this generally favorable review it seems unfortunate that most of its title so closely is that of our journal. I suspect that the publisher, rather than the author, is to be credited with the grotesque cover design. More suitable to the planet of the apes.

MAX D. ENGELHART

Herbert J. Walberg (Ed.) *Evaluating Educational Performance: A Sourcebook of Methods, Instruments, and Examples*. Berkeley, California: McCutchan Publishing Corporation, 1974. Pp. xxii + 395. \$12.00.

This collection of 19 papers addresses a variety of topics within the program evaluation/accountability realm of educational research. The editor's stated goal is to provide a practical orientation to evaluation of educational system effectiveness as a basis for policy formulation and decision making. The perspective is the macro-level of analysis with the school as subject. While the editor is the principal contributor (author or co-author of 7 chapters) his influence is even more pervasive. The majority of the contributors are/or have been affiliated with Chicago Circle, Chicago, Northwestern, Illinois, or Wisconsin and many chapters report research conducted in the Chicago public schools. Three of the chapters are reprinted and several others are based on previously delivered presentations. In chapter 1 Walberg gives a brief explanation of the purpose of the book followed by overviews of the remaining chapters.

Two-thirds of chapter 2 by Glass is devoted to an extended critique of the PMM (Popham-McNeil-Millman) method of teacher evaluation; the primary deficiency is low reliability, which also rules out standardized testing of pupils as a method of assessing teacher effectiveness. In chapter 3 Brophy outlines an intensive (over 1000 variables) investigation of teacher behaviors that may produce student achievement (as reflected in residual gain scores derived from standardized tests). Tentative results suggest that teachers' managerial skills, student attention level, and uncrowdedness of room are related to pupil achievement.

The Barclay Classroom Climate Inventory (BCCI), a computer scored, diagnostic instrument which assesses pupil deficits in eight areas, is described by its author in chapter 4. In chapter 5 Nielson and Kirk provide brief descriptions of five observation instruments and five self-report questionnaires for assessing classroom climates. Their review of relevant studies revealed no clear-cut relationship between climates and student achievement. This conclusion is seemingly contradicted by Anderson and Walberg in the following chapter. They concluded that the Learning Environment Inventory (a 15 scale, self-report measure of the "high inference" variety) accounted for substantial variance in measures of student learning. Are classroom climates and learning environments really different, or is the LEI simply a better instrument than the others?

In Chapter 7 Johnson describes the Minnesota School Affect Assessment (MSAA), an instrument designed to evaluate the effectiveness of schools in meeting educational objectives in the affective domain. Chapter 8 by Welch and Walberg, which is reprinted from *AERJ*, is a summary of a nation-wide, experimental evaluation of Project Physics. While there were no differences on the cognitive criteria, Project Physics students did report more favorable reactions to the non-cognitive aspects of the course. In another reprinted chapter, Eash describes an instrument for the assessment of instructional materials; reliability was low in the field test however.

Van Hove, Coleman, Rabben, and Karweit summarize their analyses of routinely published achievement test data and two indices of racial integration for elementary school children in six large American cities in chapter 10. They did not find a consistent integration effect, although there were specific effects within cities and grade levels. Chapter 11 by Jensen (reprinted) presents a detailed report of his investigation of the determinants of white/nonwhite differences in scholastic achievement in a California elementary school district. When the ethnic-racial groups were statistically equated for background and ability variables, differences in achievement disappeared.

Chapters 12 through 16 summarize the results of a series of studies conducted in the Chicago public schools by Walberg and his associates. Walberg and Bagen report a limited study of educational equality in chapter 12. While attendance patterns were highly segregated with minority schools evidencing lower achievement with inferior teachers, expenditures were fairly equal across schools. In the next chapter the same authors present the results of multiple regression analyses to explain variability in reading achievement at four grade levels. Not unexpectedly, earlier achievement accounted for most of the variance in later achievement. The regression equations were used to select two high- and two low-achieving elementary schools. In chapter 14 Talmage and Rippey discuss the results of their



observations of these four schools: their findings were generally not consistent with previous studies. Powell and Eash report a similar study of two high- and two low-achieving high schools in chapter 15. While the anecdotal descriptions in both chapters are interesting, the methodology is suspect on two counts: (1) the criterion (adjusted gain in reading achievement) is a statistical extrapolation, and (2) the samples are too small for anything except a pilot study. The last Chicago study by Coughlan and Cooke used the same methodology as the previous studies. A work attitudes survey was completed by teachers in high- and low-performance elementary schools and the results were discussed.

The next three chapters outline recently developed graphical methods which may be used to summarize geographically-related data. In chapter 17 McIsaac reviews trend surface analysis and gives three examples of applications. Spuck provides an overview of geocode analysis, a procedure for summarizing student data. Both trend surface and geocode analysis summarize data in the form of "contour maps." In chapter 19 Walberg and Borgen describe a regression analysis of the Chicago elementary school data using three spatial models (concentric, status, and sector) to explain differences among schools.

The final chapter by Walberg (a 1972 AAAS presentation) is an interesting, if not somewhat rambling, review of a variety of issues which relate to the future of education, e.g., genetic versus environmental explanations of behavior, increased intelligence in the population, social and economic correlates of intellectual development, effects of education on basic abilities, etc. He concludes the chapter by questioning the reliance on standardized tests as the most important measures of educational outcomes, suggesting that they "probably do far more harm than good." Ironically, Walberg ends the book expressing a viewpoint that undermines the validity of the principal criterion employed in many of the previous 19 chapters.

In summary, the volume contains a variety of review articles and research studies which summarize instruments and findings and illustrate various research strategies and methodologies in educational program evaluation. The collection is uneven in style and quality, but this is unavoidable with contributed books. The value of the volume as a sourcebook is reduced due to the absence of a subject index.

BRIAN BOLTON  
*University of Arkansas*















EDUCATIONAL and  
PSYCHOLOGICAL



MEASUREMENT

W. SCOTT GEHMAN, *Editor*

GERALDINE R. THOMAS, *Managing Editor*

WILLIAM B. MICHAEL, *Editor, Validity Studies and Computer Programs*

JOAN J. MICHAEL, *Assistant Editor, Validity Studies and Computer Programs*

MAX D. ENGELHART, *Book Review Editor*

LEWIS R. AIKEN, JR., *Assistant Book Review Editor*

FREDERIC KUDER, *Editor Emeritus*

#### BOARD OF COOPERATING EDITORS

ROTHY C. ADKINS, *University of Hawaii*  
LEWIS R. AIKEN, JR., *University of Illinois*  
GEOFF P. BECHTOLDT, *The University of Iowa*  
WILLIAM V. CLEMANS, *American Institutes for Research*

LEWIS D. COHEN, *University of Florida*  
ANTHONY J. CONGER, *Duke University*  
LEWIS A. DAVIS, *Research Triangle Institute*  
GEOFF A. EDGERTON, *Performance Research, Inc.*

LEWIS V. GLASS, *University of Colorado*  
P. GUILFORD, *University of Southern California, Los Angeles*

LEWIS A. HORNADAY, *Babson College*  
LEWIS E. HORROCKS, *The Ohio State University*  
LEWIS J. HOYT, *University of Minnesota*  
LEWIS D. JACOBSON, *University of Virginia*  
LEWIS C. JOHNSON II, *Jackson State University*

LEWIS G. KATZENMEYER, *Duke University*  
LEWIS E. LANA, *Temple University*  
LEWIS M. LORD, *Educational Testing Service*  
LEWIS LUBIN, *Navy Medical Neuropsychiatric Research Unit, San Diego*

LEWIS L. MCQUITTY, *University of Miami, Coral Gables*

HOWARD G. MILLER, *North Carolina State University at Raleigh*

ROBERT L. MORGAN, *North Carolina State University at Raleigh*

HENRY MOUGHAMIAN, *City Colleges of Chicago*

DAVID NOVAK, *The Neuse Clinic, New Bern, N. C.*

ELIS B. PAGE, *University of Connecticut*  
NAMBURY S. RAJL, *Science Research Associates, Inc.*

BEN H. ROMINE, JR., *University of North Carolina at Charlotte*

THELMA G. THURSTONE, *University of Montana*

WILLARD G. WARRINGTON, *Michigan State University*

JOHN L. WASIK, *North Carolina State University at Raleigh*

KINNARD WHITE, *University of North Carolina at Chapel Hill*

JOHN E. WILLIAMS, *Wake Forest University*

E. G. WILLIAMSON, *University of Minnesota*

VOLUME THIRTY-FIVE, NUMBER THREE, AUTUMN 1975

Diary No. 997  
Date 3/1/76



## RANDOM VARIABLES AND CORRELATIONAL OVERKILL

JOSEPH T. KUNCE AND DANIEL W. COOK  
University of Missouri-Columbia

DOUGLAS E. MILLER  
Michigan State University

Research findings may be more publishable if significant results are reported. This type of publication bias would increase the likelihood of "chance" relationships being disseminated. The implications of these assumptions are empirically investigated in a correlational analogue study. A large number of significant relationships were found in several groups of subjects between their actual scores on 45 SVIB scales and scores on 10 "experimental" scales which were determined by a set of random numbers. Furthermore, "logical" factors were shown to underly relationships which existed among scores on a given random scale with its significant correlations to SVIB scales. Considerations in such overkill in simple correlational studies are the subject-to-variable ratio, variable independence, and more stringent probability levels.

How much significance should one place in published "significant findings?" Bakan (1967), Bozarth and Roberts (1972), and Cohen (1962) have identified a publication bias that favors acceptance of research studies showing significant findings almost exclusively over nonsignificant results. Such an occurrence may create a situation where journals are "replete with research reports that have resulted in significant findings by chance factors alone (Way and Larrimore, 1973, p. 362)." Another factor affecting significant findings is statistical overkill (Kunce, 1971). Here the use of multivariate procedures (e.g., multiple correlations, factor analyses) and a subject-variable ratio less than 10 to 1 produces results that can be anticipated not to generalize

to other similar populations (Miller and Kunce, 1973). The purpose of the present investigation is to design an analogue study to investigate the incidence and implications of chance correlations upon publication bias and "overkill" in a simple, zero-order correlational design.

### *Method*

#### *Subjects*

Strong Vocational Interest Blank test scores (Form T-399) were available from a follow-up study of graduates from the University of Missouri-Columbia (Kunce, Dolliver, and Irwin, 1972). Of these subjects, 163 had participated fully in the research and 57 had not. Complete SVIB data obtained in college, however, were available for all 220 subjects on 45 of the SVIB scales. The 220 subjects were divided into four groups of 55 each. Group A consisted of 55 who had not fully participated in the follow-up study and Groups B, C, and D each had 55 subjects randomly assigned from the remaining 165.

#### *Procedure*

Each subject's data record consisted of his 45 SVIB scale scores plus scores from 10 new "experimental" scales. The scores of these experimental scales were, in reality, values extracted from a table of random numbers. Therefore, each subject had scores on a total of 55 variables.

All of the subjects' random scores on experimental scale #1 were correlated with their actual scores on each of the 45 SVIB scales for a total of 45 correlations. This procedure was repeated for each of the remaining nine experimental scales with the 45 SVIB scales. The 450 resulting correlations were computed separately for Groups, A, B, C, and D and for all groups combined. For the separate groups the ratio of subjects to variables were 1:1 (55 to 55) and the combined group 4:1 (220 to 55).

### *Results*

The number of correlations significant at or above the .05 level (2-tailed test) obtained for Groups A, B, C, and D and for the combined groups are presented in Table 1. Altogether, 17 of the intercorrelations for the 10 experimental scales with the 45 SVIB scales were significant for Group A; 26 for B; 34 for C; and 42 for D. Therefore, Groups B, C, and D had a total number of significant correlations higher than that expected by chance (i.e., assuming 23, or 5%, of the 450 intercorrela-

TABLE 1

*Frequency of Significant Correlations for Each Random Scale with 45 SVIB Scales*

Random scale	A (N = 55)	B (N = 55)	Group C (N = 55)	D (N = 55)	All (N = 220)
1	0	1	0	0	0
2	4	1	0	12	3
3	4	4	3	11	4
4	2	2	0	8	7
5	4	0	8	7	1
6	0	1	6	2	3
7	0	7	0	1	0
8	0	3	0	0	1
9	1	0	6	0	2
10	2	7	11	1	1
TOTALS	17	26	34	42	22

tions per group would be significant). The number of significant correlations ( $n = 22$ ) obtained for the four combined groups was essentially identical to that expected using a 5% chance significance rate.

### Discussion

Assume that these data from Groups A, B, C, and D represent actual findings from each of four researchers who *independently* had evaluated real, not random, scores on 10 experimental scales. Which one of the researchers, then, would have the best chances of having his results accepted for publication in a journal? The results of the findings for Groups C and D, and to a lesser extent B, would appear to have the greatest chance of being considered favorably because of the relatively high number of significant correlations. The acceptance would depend, additionally, upon the "logicalness" of the relationships reported. To explore this issue, the statistically significant intercorrelations obtained between scores on experimental scale #10 and the SVIB scales for Group C (see Table 1) were arbitrarily examined. These correlations were as follows:

Architect	+.26	
Dentist	+.29	
Physicist	+.29	
Farmer	+.31	
Personnel Director		-.40
Public Administrator		-.38
Social Science Teacher		-.33
City School		
Superintendent		-.27



Minister	-.28
Mortician	-.28
Life Insurance	-.30

From the nature of these relationships one could infer that experimental scale #10 represents a "things-people" dimension. Subsequent analysis of other random scales showed similar dimensions.

The seemingly logical relationship of scores on random scales with SVIB scales may be a consequence of the fact that the SVIB scales themselves are not independent of each other. In the above example the majority of the correlations between architect, dentist, physicist, and farmer with the other scales could be expected to be negative. Other research studies using similar measures that lack scale independence could, likewise, generate a "multiplication effect" of chance relationships.

Ironically, the results of our hypothetical researchers C and D would appear to have the best chance of getting published promulgating "false findings;" whereas, the data from Group A which may contain "truer" results (i.e., within the 5% chance expectancy range) is not likely to get published. The probability of false results being generated and disseminated would likely be reduced if a higher subject/variable ratio is used. For example, for the combined group of subjects having a subject/variable ratio of 220:55, the number of statistically significant correlations,  $N = 22$ , did not exceed that expected by chance alone. Even here caution should be exerted with regards to individual scales and their "validity." Experimental scale #4 (see Table 1) for the combined groups correlated significantly with seven of the 45 SVIB scales. This could be interpreted as three times the chance level, assuming that 5% of 45 (or approximately 2) correlations could be anticipated to be significant.

The results of this study support the position that many studies accepted for publication may have largely chance findings in spite of their reported statistical significance. The face validity of chance significant findings resulting from scale interdependence further masks the true relationship. Several considerations may reduce premature acceptance of chance findings in simple correlational studies:

1. Increase the number of subjects in relationship to the number of variables to at least 10:1 as in multivariate studies.
2. Use more stringent probability levels when low subject/variable ratios are unavoidable.
3. Use *independent* validity generalization samples before publishing data based on low subject/variable ratios.

We are not in a position, nor have the expertise, to determine how many significant correlations are needed in a correlational table to be significantly greater than chance when scales are not independent. Undoubtedly, this determination could be a complex procedure and would vary according to the degree of interscale independence. The computational complexities involved or the rigid interpretation of guidelines could lead to suppression of new and potentially useful psychological findings.

Greater appreciation and awareness of the realities of statistical and correlational overkill, rather than complex mathematics, may be sufficient to reduce premature acceptance of superficially impressive findings. This conclusion is exemplified by the recommendations of other investigators. Tversky and Kahneman (1971) cautioned against the expectation that findings from small samples will generalize to other small replication samples even if both are thought to be representative of the population. Furthermore, if replication samples are smaller than the developmental sample, the power of the test is diminished which reduces the chance of gaining significant results. Lykken (1968) believes that significance level is the least important attribute of a good experiment and is not sufficient for indicating whether a theory has been corroborated, an empirical fact confidentially established, or whether a study should be published. He feels all experiments ideally should be replicated before publication. And, finally, Winer (1971, p. 14) has taken the position that "no absolute standard can be set up for determining the appropriate level of significance and power that a test should have. The level of significance used in making statistical tests should be gauged in part by the power of practically important alternative hypotheses at varying levels of significance."

## REFERENCES

- Bakan, D. The test of significance. In D. Bakan (Ed.) *On method: Toward a reconstruction of psychological investigation*. San Francisco: Jossey-Bass, Inc., 1967.
- Bozarth, J. E. and Roberts, R. R. Signifying significant significances. *American Psychologist*, 1972, 27, 774-775.
- Cohen, J. The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 1962, 65, 145-153.
- Kunce, J. T. Prediction and statistical overkill. *Measurement and Evaluation in Guidance*, 1971, 4, 38-42.
- Kunce, J. T., Dolliver, R., and Irvin, J. Perspectives on interpreting the validity of the SVIB-M. *Vocational Guidance Quarterly*, 1972, 21, 36-42.

- Lykken, D. T. Statistical significance in psychological research. *Psychological Bulletin*, 1968, 70, 151-159.
- Miller, D. E. and Kuncce, J. T. Statistical overkill revisited. *Measurement and Evaluation in Guidance*, 1973, 6, 157-163.
- Tversky, A. and Kahneman, D. Belief in the law of small numbers. *Psychological Bulletin*, 1971, 76, 105-110.
- Way, J. G. and Larrimore, D. L. A joinder on signifying significant significance. *American Psychologist*, 1973, 28, 361-362.
- Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1971.

## INDEPENDENCE PROBLEMS FOR CERTAIN TESTS BASED ON THE SHINE-BOWER ERROR TERM

LESTER C. SHINE II  
Texas A & M University

It is shown for the Shine-Bower single-subject ANOVA that the numerator and denominator of all  $F$  tests based on the Shine-Bower error term are independent of each other. It is also shown that the same property holds for all such tests in the Shine Combined ANOVA except for the test for the trial by subject interaction.

THIS article is intended to clarify the independence of the numerator and denominator of certain  $F$  tests which use the Shine-Bower error term ( $MSE'$ ) as a denominator. These tests occur in the Shine-Bower single-subject ANOVA (Shine and Bower, 1971; Shine, 1973b) and in the Shine Combined ANOVA (Shine, 1973a, 1974). It has been indicated by Shine (1974, footnote on p. 50) that a lack of independence occurs only in the Combined ANOVA test for an interaction between trials and subjects.

### *Shine-Bower Single-Subject ANOVA (General Case)*

$MSE'$  is used in this fixed factor design to test all sources of variation except the main effect of trials. It is assumed for  $MSE'$  that the main effect of trials changes slowly across trials. Independent additive contributions to  $MSE'$  are obtained by squaring and summing every other successive difference between trial means (an even number of trials is assumed), after collapsing across all other factors. These successive differences themselves may be expressed as successive differences between the standard linear forms representing trial main effects (Scheffé, 1959). This set of trial main effect linear forms is orthogonal to any of the sets of standard linear forms on which mean squares for

the other sources of variation in the design are based (Scheffé, 1959). Consequently, it follows immediately from the normality and independence assumptions of the Shine-Bower ANOVA that  $MSE'$  is statistically independent of all mean squares in the design except for the trial main effect mean square.

### *Shine Combined ANOVA (General Case)*

This design is effectively a standard repeated measures design in which  $MSE'$  is used only for testing subject sources of variation.  $MSE'$  is obtained by pooling across suitable subjects the corresponding sums of squares and degrees of freedom associated with each individual subject's Shine-Bower error term.  $MSE'$  is clearly a function of changes across trials within subjects and within the nonrepeated factors, after collapsing across all repeated factors. Thus,  $MSE'$  must be a function of the standard linear forms for the main effect of trials, for all interactions involving only trials and one or more of the nonrepeated factors, and for the nested trial by subject interaction (the subject factor is nested in the nonrepeated factors). In the sense stated above, these forms are orthogonal to all other sources of variation in the design (Scheffé, 1959). Consequently, under the normality assumption and under the usual homogeneity assumptions for variances and covariances (Winer, 1971),  $MSE'$  is statistically independent of all subject effect mean squares except for the nested trial by subject interaction mean square.

### *Conclusion*

Only the  $F$  test in the Combined ANOVA for testing the nested trial by subject interaction (non-nested for the special case of a completely repeated design) presents a problem regarding the independence of the numerator and denominator ( $MSE'$ ). This conclusion holds whether or not any null hypotheses are true and whether or not the slow change assumption for  $MSE'$  is met. In any case, the  $F$  test for the trial by subject interaction must be regarded as very approximate in nature. The present author is currently investigating the exact distribution of this statistic.

### REFERENCES

- Scheffé, H. *The analysis of variance*. New York: John Wiley & Sons, Inc., 1959.
- Shine, L. C. A design combining the single-subject and multi-subject



- approaches to research. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1973, 33, 763-766. (a)
- Shine, L. C. A multi-way analysis of variance for single-subject designs. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1973, 33, 633-636. (b)
- Shine, L. C. An extension of the Shine combined analysis of variance. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1974, 34, 47-52.
- Shine, L. C. and Bower, S. M. A one-way analysis of variance for single-subject designs. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1971, 31, 105-113.
- Winer, B. J. *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill, 1971.



## ON THE INDEPENDENCE OF VARIABLE SETS

CHARLES D. DZIUBAN, EDWIN C. SHIRKEY, AND THOMAS O. PEEPLES  
Florida Technological University

An illustration of a test for independence was provided with a mixed set of variables. The matrix consisted of 10 tests of interest and four random deviates in which the relationship between sets was demonstrated to be minimal. The result was discussed for a situation in which factoring methods might be considered.

RECENTLY, Shaycoft (1970) presented a clear example in which results yielded by the method of principal components would lead to an erroneous interpretation. The procedure was applied to a "mixed" set of variables—ten measures of interest from project TALENT and four random deviates ( $N = 3689$ ). The obtained results would have forced one to attempt an interpretation of random variables as the basis of a meaningful component.

Dziuban and Harris (1973) illustrated that such a result would be guarded against by application of the image and factor analytic models. They noted that the random variables did not warrant any interpretation when those procedures were applied to the "mixed set." Their recommendations included abandonment of principal components in favor of factor analytic and image methods when one has reason to believe that some variables in a set are essentially random.

In a case such as the one presented by Shaycoft, it might be expected that the ten project TALENT and four random variables, the two sets, were independent of each other. That hypothesis might be tested using the Heck largest root distribution (Morrison, 1967). If the largest eigenvalue of the matrix  $R_{11}^{-1}R_{12}R_{22}^{-1}R_{12}'$  exceeds the 100  $\alpha$  percentage point of the distribution, the independence hypothesis may be rejected.

In this case  $R_{11}$  comprised the correlations among the TALENT

TABLE 1  
The Matrix  $R_{11}^{-1}R_{12}R_{22}^{-1}R_{12}$

1	2	3	4	5
$-5.2 \times 10^{-5}$	$1.9 \times 10^{-4}$	$6.1 \times 10^{-4}$	$2.7 \times 10^{-4}$	$-4.0 \times 10^{-4}$
$-1.0 \times 10^{-3}$	$1.1 \times 10^{-3}$	$1.5 \times 10^{-3}$	$-5.1 \times 10^{-4}$	$-4.3 \times 10^{-4}$
$-5.5 \times 10^{-4}$	$6.4 \times 10^{-4}$	$1.0 \times 10^{-3}$	$-2.6 \times 10^{-4}$	$-3.8 \times 10^{-4}$
$1.2 \times 10^{-4}$	$1.5 \times 10^{-3}$	$5.0 \times 10^{-4}$	$8.0 \times 10^{-4}$	$-4.4 \times 10^{-4}$
$2.5 \times 10^{-3}$	$2.6 \times 10^{-3}$	$8.2 \times 10^{-3}$	$7.4 \times 10^{-3}$	$-6.8 \times 10^{-3}$
$-6.1 \times 10^{-4}$	$5.0 \times 10^{-4}$	$1.1 \times 10^{-3}$	$-7.5 \times 10^{-4}$	$-4.4 \times 10^{-4}$
$-1.2 \times 10^{-3}$	$1.1 \times 10^{-3}$	$1.6 \times 10^{-3}$	$-9.9 \times 10^{-4}$	$-4.5 \times 10^{-4}$
$-3.7 \times 10^{-4}$	$3.2 \times 10^{-4}$	$6.1 \times 10^{-4}$	$-5.7 \times 10^{-4}$	$-1.9 \times 10^{-4}$
$-5.5 \times 10^{-4}$	$4.1 \times 10^{-4}$	$4.4 \times 10^{-4}$	$-1.0 \times 10^{-3}$	$3.8 \times 10^{-3}$
$-2.2 \times 10^{-4}$	$2.7 \times 10^{-4}$	$5.2 \times 10^{-4}$	$-3.3 \times 10^{-4}$	$-2.1 \times 10^{-4}$
6	7	8	9	10
$9.0 \times 10^{-5}$	$3.3 \times 10^{-4}$	$5.2 \times 10^{-4}$	$-5.5 \times 10^{-4}$	$2.0 \times 10^{-4}$
$1.8 \times 10^{-4}$	$1.5 \times 10^{-3}$	$6.2 \times 10^{-4}$	$-6.5 \times 10^{-4}$	$-6.3 \times 10^{-4}$
$-2.2 \times 10^{-4}$	$6.6 \times 10^{-4}$	$7.3 \times 10^{-4}$	$-3.2 \times 10^{-4}$	$7.0 \times 10^{-4}$
$-2.9 \times 10^{-4}$	$1.6 \times 10^{-3}$	$5.1 \times 10^{-4}$	$-6.1 \times 10^{-4}$	$5.2 \times 10^{-4}$
$-2.2 \times 10^{-4}$	$-1.1 \times 10^{-4}$	$1.7 \times 10^{-4}$	$1.4 \times 10^{-3}$	$5.3 \times 10^{-4}$
$8.3 \times 10^{-4}$	$1.2 \times 10^{-3}$	$7.0 \times 10^{-4}$	$-6.0 \times 10^{-4}$	$-5.0 \times 10^{-4}$
$5.4 \times 10^{-4}$	$1.8 \times 10^{-3}$	$7.3 \times 10^{-4}$	$-6.3 \times 10^{-4}$	$-7.8 \times 10^{-4}$
$-7.0 \times 10^{-5}$	$-3.9 \times 10^{-4}$	$5.6 \times 10^{-4}$	$1.4 \times 10^{-3}$	$8.4 \times 10^{-4}$
$-3.8 \times 10^{-4}$	$2.6 \times 10^{-4}$	$4.2 \times 10^{-4}$	$5.0 \times 10^{-4}$	$1.2 \times 10^{-3}$
$-5.2 \times 10^{-4}$	$3.0 \times 10^{-3}$	$6.8 \times 10^{-4}$	$1.2 \times 10^{-4}$	$1.6 \times 10^{-3}$

variables,  $R_{22}$  those among the random variables, and  $R_{12}$  their inter-correlations. Referring to the largest root distribution, the value needed to reject the hypothesis  $\alpha (= .01)$  is approximately .025. The largest obtained eigenvalue of the matrix was .0035 so that the independence hypothesis would not be rejected.<sup>1</sup> This result suggests that if one has reason to suspect that some variables in a set are *unrelated* to the domain of interest, assessment might be made prior to any "factoring" procedures.

## REFERENCES

- DZIUBAN, D. AND HARRIS, C. W. On the extraction of components and the applicability of the factor model. *American Educational Research Journal*, 1973, 10, 93-99.
- Morrison, D. F. *Multivariate statistical methods*. New York: McGraw-Hill Book Co., 1967.
- Shaycoft, M. F. The eigenvalue myth and the data reduction fallacy. (Paper presented at the annual meeting of The American Educational Research Association, Minneapolis, Minnesota, March, 1970).

<sup>1</sup> Largest canonical correlation = .059 ( $x^2 = 31.96$ , D.F. = 40).

## SAMPLING CHARACTERISTICS OF KELLEY'S $\epsilon^2$ AND HAYS' $\hat{\omega}^2$

ROBERT M. CARROLL

U.S. Army Research Institute for the Behavioral and Social Sciences

LENA A. NORDHOLM

Indiana University Northwest<sup>1</sup>

Statistics used to estimate the population correlation ratio were reviewed and evaluated. The sampling distributions of Kelley's  $\epsilon^2$  and Hays'  $\hat{\omega}^2$  were studied empirically by computer simulation within the context of a three level one-way fixed effects analysis of variance design. These statistics were found to have rather large standard errors when small samples were used. As with other correlation indices, large samples are recommended for accuracy of estimation. Both  $\epsilon^2$  and  $\omega^2$  were found to be negligibly biased. Heterogeneity of variances had negligible effects on the estimates under conditions of proportional representativeness of sample sizes with respect to their population counterparts, but combinations of heterogeneity of variance and unrepresentative sample sizes yielded especially poor estimates.

IN spite of the consistent emphasis on  $p$ -levels by editors of psychological journals, some researchers have concerned themselves with the question of the strength of relationship between the independent and dependent variables in comparative experiments. This concern is justified and should be encouraged, since the emphasis on  $p$ -levels alone may lead to exploitation in the form of reporting trivial effects due to large sample sizes. It is questionable, however, whether the need for "practical significance" is best served by estimates of the strength

---

<sup>1</sup> This study was supported by a grant for computer time from the Computer Science Center, University of Maryland. An earlier draft of this paper was read at the Midwestern Psychological Association Meeting at Cleveland, Ohio in May 1972. This study was conducted while the authors were at the University of Maryland.



of the relationship such as the correlation ratio computed on a sample or some of the other estimators of the population correlation ratio as given by Pearson (1923), Kelley (1935), or Hays (1963).

Originally, a measure known as the correlation ratio or  $\eta^2$  (eta squared) was developed to give a descriptive index of the total relationship in a single factor fixed effect design, in contrast to  $r^2$ , which only indicates linear relationship. In the analysis of variance notation  $\eta^2$  was defined as:

$$\eta^2 = \frac{SS_T - SS_W}{SS_T} = \frac{SS_B}{SS_T} \quad (1)$$

where:

$SS_T$  = total sum of squares.

$SS_W$  = sum of squares within groups.

$SS_B$  = sum of squares between groups.

The population value of  $\eta^2$  was similarly defined by substituting population sums of squares for their sample counterparts, which lead to the formula:

$$\eta^2_{\text{pop}} = \frac{\sigma_y^2 - \sigma_e^2}{\sigma_y^2} \quad (2)$$

where:

$\sigma_y^2$  = the variance of the dependent variable.

$\sigma_e^2$  = the common homogeneous variance of the  $Y_{ij}$  about  $\mu_j$ .

However, recognizing that  $\eta^2$  (Formula 1) is not an unbiased estimate of the correlation ratio in the population, Pearson (1923) suggested an improved (less biased) large sample approximation of the correlation ratio in the population. His estimate was:

$$\eta_p^2 = \frac{\eta_o^2 - (J - 3)/N}{1 - (J - 3)/N} \quad (3)$$

$\eta_p^2$  = estimate of population correlation ratio.

$\eta_o^2$  = observed correlation ratio on sample.

$J$  = number of data arrays.

$N$  = total sample size.

Kelley (1935) proposed another estimate of the population correlation ratio, which he believed to be unbiased. This statistic, which he called  $\epsilon^2$  (epsilon squared) was formulated by substituting what Kelley thought were unbiased estimators for  $\sigma_y^2$  and  $\sigma_e^2$  in Equation 2. His estimate of the population correlation ratio so determined was:

$$\epsilon^2 = \frac{SS_T/(N - 1) - SS_W/(N - J)}{SS_T/(N - 1)} \quad (4)$$

where:

$N$  = total number of observations taken.

$J$  = number of data arrays.

Glass and Hakstian (1969) pointed out that epsilon squared is actually not an unbiased estimate of the population correlation ratio since the ratio of unbiased estimators is not generally an unbiased estimate of the ratio itself (see Olkin and Pratt, [1958]).

Hays (1963) derived still another measure of the strength of relationship between a categorical variable  $X$  and the dependent variable  $Y$ . This index was called  $\omega^2$  (omega squared) and was defined for the population as:

$$\omega^2 = \frac{\sigma_y^2 - \sigma_{y|x}^2}{\sigma_y^2} \quad (5)$$

where:

$\sigma_y^2$  = marginal variance of  $Y$ .

$\sigma_{y|x}^2$  = conditional variance of  $Y$  given any  $X$ .

This is equivalent to the definition of the correlation ratio defined for a population. According to Hays (1963) a "rough" estimate of the population correlation ratio is given by:

$$\hat{\omega}^2 = \frac{SS_B - (J - 1)MS_W}{SS_T + MS_W} \quad (6)$$

where:

$SS_T$  = sum of squares total.

$SS_B$  = sum of squares between groups.

$MS_W$  = mean squares within groups.

$J$  = number of groups.

Glass and Hakstian (1969) dealt with the relationship between  $\epsilon^2$  and  $\hat{\omega}^2$  and showed that epsilon squared could be written as:

$$\epsilon^2 = \frac{SS_B - (J - 1)MS_W}{SS_T} \quad (7)$$

The difference between Formulas 6 and 7 can be seen in the denominators. Glass and Hakstian suggested that this difference is due to the varying definitions of  $\sigma_y^2$  that Hays and Kelley employ. We contend that the definition of  $\sigma_y^2$  is not equivocal, and that Equations 6 and 7 differ due to the different estimates of  $\sigma_y^2$  used by Hays and Kelley. Glass and Hakstian pointed out that  $E[SS_T/(N - 1)] = \sigma_e^2 + \sum n_j \alpha_j^2 / (N - 1)$  where  $N$ ,  $n_j$  are sample values and  $\alpha_j$  is the effect of treatment  $j$ . Hays defined  $\sigma_y^2 = \sigma_e^2 + \sum n_j \alpha_j^2 / N$  in a single factor fixed effect design but under the restriction that the relative sample sizes in

the various treatment groups be equal to the probability of randomly observing a case from the corresponding population. Consequently Hays' definition of  $\sigma_y^2$  is not dependent upon sample size as he requires  $n_j/N$  to be fixed. However  $E[SS_T/(N-1)]$  does depend on sample size as  $n_j/(N-1)$  will vary with different sample sizes even when the proportional representativeness of the samples equals its population counterpart. One obviously cannot define a population parameter whose value is dependent upon the sample size used to estimate it. Therefore,  $\sigma_y^2$  cannot be appropriately defined as  $\sigma_e^2 + \sum n_j \alpha_j^2 / (N-1)$ . Hays' definition is the only one appropriate and consequently  $SS_T/(N-1)$  is a biased estimate of  $\sigma_y^2$  within the context of a one-way fixed effects model. Hays developed his definition of  $\sigma_y^2$  by assuming the following fixed effect model:

$$Y_{ij} = \mu + \alpha_j + e_{ij} \quad (8)$$

where  $\sum n_j \alpha_j = 0$  with the  $e_{ij}$  independently and normally distributed with expectation zero and variance  $\sigma_e^2$ . By definition  $\sigma_y^2 = E(Y_{ij} - \mu)^2 = E(\alpha_j + e_{ij})^2 = E(\alpha_j^2) + E(e_{ij}^2) = \sigma_e^2 + E(\alpha_j^2)$ . The term  $\alpha_j$  is a discrete random variable with probability  $n_j/N$  of taking on the value  $\alpha_j$  when the sample proportions are representative of the population so  $E(\alpha_j^2)$  would be given by  $\sum n_j \alpha_j^2 / N$  resulting in  $\sigma_y^2 = \sigma_e^2 + \sum n_j \alpha_j^2 / N$ .

Glass and Hakstian showed that the two estimates,  $\epsilon^2$  and  $\hat{\omega}^2$ , of the population relationship would be essentially the same in practice, since their relationship can be written as:

$$\epsilon^2 = \hat{\omega}^2 + \frac{MS_W \hat{\omega}^2}{SS_T} \quad (9)$$

It can be readily seen as the sample size increases or as the error variance decreases, the estimates converge.

The popularity of Hays'  $\hat{\omega}^2$  as opposed to the lack of emphasis given to Kelley's  $\epsilon^2$  over the years in psychological journals is difficult to explain. Glass and Hakstian discussed some problems with the interpretation of  $\hat{\omega}^2$  (and by implication  $\epsilon^2$ ), which they cited as the probable reason for the lack of popularity of  $\epsilon^2$  over the years. Their main argument was that these measures depend too much on the specific levels chosen for the independent variable. They postulated that although research workers claim to be dealing with a fixed effects model they are actually concerned with their variables as molar constructs and not with the levels actually administered in the experiment. It seems to the present authors that although it is frequently true the fixed effects model is used in situations not satisfying all requirements for such a model, there are circumstances where the fixed effects model is appropriate. Yet, Glass and Hakstian's warning should be heeded

that such measures can be very misleading when used to estimate the strength of association for variables where not all levels of interest are included in the design.

It was the purpose of this study to empirically investigate some of the characteristics of the sampling distributions of  $\epsilon^2$  and  $\hat{\omega}^2$  under a limited set of conditions. The procedure consisted of taking random samples from populations with known correlation ratios and comparing the estimates so obtained with the true population values. It is also hoped that the results will give additional impetus to the discussion of the appropriateness of measures of association in the context of fixed effects analysis of variance.

### *Procedure*

#### *Parameters Relevant to Generation of Treatment Populations*

Since the study was concerned with measures of association in the fixed effects analysis of variance context, the first consideration was the number of independent variables and the number of levels chosen for such variables. It was decided to carry out all experiments within the framework of a three treatment level one-way analysis of variance design in order to keep matters simple and also because there appears to be a direct generalization to higher order designs. Five levels of the population correlation ratio were used:  $\eta^2 = .00$ ,  $\eta^2 = .05$ ,  $\eta^2 = .15$ ,  $\eta^2 = .40$ , and  $\eta^2 = .75$ .

Both Kelley and Hays made use of the assumption of equal within treatment population variances in the development of their formulas for  $\epsilon^2$  and  $\hat{\omega}^2$ . The homogeneity of variance assumption has been shown not to be of critical importance in the  $F$  test, provided equal sample sizes are used (Box, 1954, Norton studies, cited in Lindquist, 1953). In empirical research this assumption is often violated, and it is therefore desirable that the effects of heterogeneous variances on the measures used to estimate the strength of relationship be studied. Three levels of heterogeneity of variances were used: zero heterogeneity (homogeneous variances); slight heterogeneity (ratios 3:2:1 from largest to smallest variance); and marked heterogeneity (ratios 10:4:1 from largest to smallest variance). Since experimental treatments generally affect both means and variances such that larger means are usually associated with larger variances (Norton studies, cited in Lindquist, 1953), heterogeneous variances were created accordingly.

Table 1 shows for the fixed values of the treatment means the within treatment error variances giving the desired population  $\eta^2$  values. All

TABLE I  
*Within Treatment Means and Variances Used to Generate Treatment Populations.*

Variance Ratios	Level of Fixed Factor	Population eta squared													
		.00			.05			.15			.40			.75	
		$\mu$	$\sigma_e^2$	$\mu$	$\sigma_e^2$	$\mu$	$\sigma_e^2$	$\mu$	$\sigma_e^2$	$\mu$	$\sigma_e^2$	$\mu$	$\sigma_e^2$	$\mu$	$\sigma_e^2$
1:1:1	1	150.0	64.00	140.0	1266.67	140.0	377.78	140.0	100.00	140.0	22.22				
	2	150.0	64.00	150.0	1266.67	150.0	377.78	150.0	100.00	150.0	22.22				
	3	150.0	64.00	160.0	1266.67	160.0	377.78	160.0	100.00	160.0	22.22				
1:2:3	1	150.0	32.00	140.0	633.33	140.0	188.90	140.0	50.00	140.0	11.11				
	2	150.0	64.00	150.0	1266.66	150.0	377.80	150.0	100.00	150.0	22.22				
	3	150.0	96.00	160.0	1899.99	160.0	566.70	160.0	150.00	160.0	33.33				
1:4:10	1	150.0	25.00	140.00	253.33	140.0	75.56	140.0	20.00	140.0	4.44				
	2	150.0	100.0	150.0	1013.32	150.0	302.24	150.0	80.00	150.0	17.76				
	3	150.0	250.00	160.0	2533.30	160.0	755.60	160.0	200.00	160.0	44.40				



treatment populations were normally distributed with the means and variances as shown in Table 1.

### *Parameters Relevant to the Sampling Experiments*

A series of experiments were simulated by drawing random samples from the treatment populations specified above. Total sample size was varied at three levels  $N = 15$ ,  $N = 30$ , and  $N = 90$ . Box (1954) showed that the violation of homogeneity of variance assumption had little effect on the probability of the  $F$  statistic exceeding the 5% significance point, provided equal sample sizes per treatment condition were used. However, unequal sample sizes did seriously affect the robustness of the  $F$  statistic to violations of homogeneity of variance. A similar effect should be found for the estimates of strength of association,  $\epsilon^2$  and  $\hat{\omega}^2$ , since they are directly related to  $F$ . It can be easily shown that:

$$\hat{\omega}^2 = \frac{F - 1}{F + \frac{N - J + 1}{J - 1}} \quad (10)$$

where

$N$  = total sample size.

$J$  = number of treatment levels.

and

$$\epsilon^2 = \frac{F - 1}{F + \frac{N - J}{J - 1}} \quad (11)$$

The results of Box were based on the probability of an  $F$  exceeding some point, while the present study is concerned more with the effect on the specific magnitude of the  $F$  statistic and consequently,  $\epsilon^2$  and  $\hat{\omega}^2$ . The Norton study did look at the effect of heterogeneous variances on the magnitude of the  $F$  statistic, but only for the case of equal sample sizes. In the present study for each level of total sample size, the within treatment sample sizes were either equal or in the ratio 3:5:7. For the heterogeneous variance conditions, two unequal sample size conditions were used, so that the largest sample size was associated with the largest within treatment variance in one condition, and associated with the smallest within treatment variance under the other condition (i.e., the ratio was varied from 3:5:7 to 7:5:3).

In all there were five levels of population  $\eta^2$ , three levels of heterogeneity of variance, three levels of total sample size, and three levels of equal-unequal sample sizes. Within each cell 1000 simulated experiments were carried out with  $F$ ,  $\epsilon^2$ , and  $\hat{\omega}^2$  being computed for each.

In his derivation of  $\epsilon^2$ , Kelley states that no systematic error will be introduced if the numbers in the arrays for a sample give proportions which differ in only a random manner from the population proportions. However, a word of justification for estimating  $\hat{\omega}^2$  by use of Formula 6 for the unequal sample size conditions seems warranted since this formula was developed for the situation where "... the proportional representation of cases in the  $J$  samples is the same as the proportions in the respective populations ..." (Hays, 1963, p. 382). Hays did not suggest an alternative estimate of  $\omega^2$  when the proportional representativeness of the samples fails to match its population counterpart. Vaughan and Corballis (1969) although strongly urging the use of equal sample sizes did offer a "fair approximation" appropriate provided samples were approximately equal. They developed their argument by noting that in a one-way fixed effects model,  $E(MS_B) = \sigma_e^2 + (n \sum_{j=1}^J \alpha_j^2)/(J-1)$ , where  $n$  is the number of observations per cell. In order to estimate the variance between groups,  $\sigma_B^2 = \sum \alpha_j^2/J$ , they used:

$$\hat{\sigma}_\alpha^2 = (J-1)(MS_B - MS_w)/nJ. \quad (12)$$

But for the unequal sample size case,  $E(MS_B) = \sigma_e^2 + \sum_{j=1}^J n_j \alpha_j^2/(J-1)$ , so one cannot simply divide by  $n$  as in Equation 12 to obtain an unbiased estimate of  $\sigma_B^2$ . Vaughan and Corballis instead suggested dividing by the mean  $n_j$ ,  $(\bar{n})$ . The  $\hat{\omega}^2$  then becomes:

$$\begin{aligned} \hat{\omega}^2 &= \frac{(J-1)(MS_B - MS_w)/\bar{n}J}{\{(J-1)(MS_B - MS_w) + \bar{n}JMS_w\}/\bar{n}J} \\ &= \frac{SS_B - (J-1)MS_w}{SS_B + (\bar{n}-1)JMS_w + MS_w} \\ &= \frac{SS_B - (J-1)MS_w}{SS_T + MS_w}. \end{aligned} \quad (13)$$

Thus, the estimate of  $\omega^2$  for the unequal sample size case reduces to the identical estimate, Equation 6, for the case of proportional representation.

### Results

Means and standard deviations of the obtained  $\hat{\omega}^2$  and  $\epsilon^2$  were computed for each of the 120 cells. In Table 2 the means are given as the upper numbers with the standard deviations given just below them. The means indicate that  $\hat{\omega}^2$  is slightly negatively biased ( $Z = -4.31, p < .001$ , using data from equal sample size and homogeneous variance

conditions). Any bias in  $\epsilon^2$  is not evident from the data of Table 2. The combination of homogeneous variances and unequal  $n$  yielded consistent mean underestimates across all conditions of  $\eta^2$  and total  $N$ . With heterogeneous variances and unequal  $n$ ,  $\hat{\omega}^2$  and  $\epsilon^2$  were either overestimates or underestimates depending on the particular combination of unequal  $n$  with within cell error variance. When the sample with largest  $n$  was taken from the treatment population with largest variance,  $\epsilon^2$  and  $\hat{\omega}^2$  substantially underestimated the population relationship. However, for the reverse case (largest  $n$  from treatment population with smallest variance)  $\hat{\omega}^2$  and  $\epsilon^2$  substantially overestimated the population association. It can also be seen that heterogeneity of variance had little effect on the estimates under conditions of equal  $n$ .

Hays did not provide the standard error of  $\hat{\omega}^2$ , but Kelley did derive an approximation of the standard error of  $\epsilon^2$  which he gave as:

$$\sigma_{\epsilon^2} = \frac{1 - \epsilon^2}{\sqrt{N - 1}} \left( \frac{2(J - 1)}{N - J} + 4\epsilon^2 \right)^{1/2} \quad (14)$$

According to Kelley this standard error is appropriate unless  $1/N$  is not small in comparison with  $1/\sqrt{N}$ . As with  $\epsilon^2$ , the homogeneity of variance assumption was used in deriving Formula 14. If population  $\eta^2$  is substituted for  $\epsilon^2$  in Formula 14, the resulting estimates for the conditions of homogeneous variances and equal sample sizes are in close agreement with the standard errors for  $\epsilon^2$  given in Table 2, but are consistently slightly larger.

The data of Table 2 indicate that both  $\hat{\omega}^2$  and  $\epsilon^2$  have large standard deviations when small samples are used. For instance when  $\eta^2 = .15$  or  $.40$ , the standard error of  $\hat{\omega}^2$  was consistently close to  $.20$  for  $N = 15$ , and close to  $.13$  for  $N = 30$ , and still as large as  $.07$  for  $N = 90$ . A comparison between  $\hat{\omega}^2$  and  $\epsilon^2$  favors  $\hat{\omega}^2$  as a slightly more efficient estimator than  $\epsilon^2$ , since the standard deviations of  $\hat{\omega}^2$  were consistently somewhat lower than those of  $\epsilon^2$ . Table 2 also shows that as total  $N$  increases the standard deviations of the two indices decreased. Just as the means of  $\hat{\omega}^2$  and  $\epsilon^2$  were affected by the particular combination of heterogeneity of variance and unequal  $n$ , the standard deviations were similarly affected. In fact, within each condition of population  $\eta^2$ , there was a positive relationship between the means and the standard deviations such that the largest mean was associated with the largest standard deviation. A final observation regarding the standard deviations is that the standard errors varied with the degree of the population relationship. When the association in the population was very strong ( $\eta^2 = .75$ ) or very small ( $\eta^2 = .00$ ) the standard deviations were consistently lower than in any of the other conditions. The estimates

TABLE 2  
Means and Standard Deviations of  $\hat{\omega}^2$  and  $\epsilon^2$  for Each Condition

n	.00		.05		.15		.40		.75	
	Ratio of variances		Ratio of variances		Ratio of variances		Ratio of variances		Ratio of variances	
	1:1:1	1:4:10	1:1:1	1:2:3	1:1:1	1:2:3	1:1:1	1:2:3	1:1:1	1:2:3
5, 5, 5	-.000	.012	.045	.059	.154	.135	.145	.382	.395	.389
	.140	.160	.154	.171	.198	.191	.202	.191	.196	.20
3, 5, 7	-.008	-.036	.039	.018	.111	.098	.072	.350	.395	.299
	.121	.116	.155	.151	.176	.172	.169	.203	.191	.18
7, 5, 3		.075		.081	.167	.167	.213		.406	.460
		.197		.179	.206	.206	.233		.19	.209
10, 10, 10	-.000	.001	.048	.048	.145	.149	.143	.398	.395	.388
	.069	.078	.100	.097	.121	.126	.132	.131	.130	.13
$\hat{\omega}^2$ 6, 10, 14	-.002	-.019	.042	.029	.123	.111	.095	.363	.328	.308
	.064	.054	.093	.082	.119	.111	.105	.135	.124	.126
14, 10, 6		.035		.061	.162	.162	.192		.407	.437
		.100		.111	.137	.137	.154		.136	.151
30, 30, 30	-.000	-.001	.047	.048	.148	.144	.147	.398	.401	.399
	.020	.025	.046	.045	.069	.069	.070	.074	.075	.080
18, 30, 42	-.001	-.007	.044	.032	.137	.115	.107	.370	.338	.318
	.020	.016	.045	.039	.065	.058	.056	.075	.070	.071
42, 30, 18		.008		.056	.155	.155	.181		.407	.438
		.033		.053	.076	.076	.089		.081	.08
					.065	.065			.746	.750
					.050	.050			.036	.038
					.032	.032			.725	.700
					.036	.036			.037	.042
					.066	.066			.757	.779
					.028	.028			.039	.041

5, 5, 5	-.002	-.004	.011	.046	.061	.048	.161	.141	.151	.395	.409	.402	.755	.756	.760
	.147	.148	.169	.162	.180	.173	.206	.199	.210	.194	.199	.212	.101	.105	.105
3, 5, 7	-.010	-.029	-.039	.040	.018	-.004	.116	.102	.075	.363	.331	.312	.731	.702	.689
	.134	.130	.123	.163	.159	.150	.184	.180	.177	.208	.196	.193	.110	.115	.122
7, 5, 3		.029	.077		.084	.114	.173	.173	.221		.420	.474		.755	.800
		.161	.206		.187	.217	.215	.240			.202	.211		.101	.105
10, 10, 10	-.000	-.000	.001	.049	.049	.046	.149	.153	.146	.405	.403	.395	.753	.754	.749
	.071	.070	.080	.103	.099	.104	.124	.129	.135	.132	.131	.137	.065	.066	.076
$\epsilon^2$ 6, 10, 14	-.002	-.013	-.020	.043	.030	.015	.126	.114	.098	.371	.335	.315	.731	.702	.685
	.066	.058	.056	.096	.084	.080	.122	.113	.108	.136	.125	.127	.073	.073	.080
14, 10, 6		.014	.036		.062	.086	.166	.196	.196		.414	.445		.761	.792
		.083	.103		.114	.132	.140	.157			.137	.151		.067	.072
30, 30, 30	-.000	.000	-.001	.048	.048	.051	.149	.145	.148	.400	.404	.401	.749	.752	.750
	.020	.022	.025	.046	.046	.050	.070	.069	.071	.074	.075	.080	.036	.038	.042
18, 30, 42	-.001	-.003	-.007	.044	.033	.030	.138	.117	.108	.373	.340	.321	.727	.702	.684
	.020	.019	.017	.045	.039	.037	.066	.059	.056	.075	.070	.071	.037	.042	.046
42, 30, 18		.005	.008		.057	.067	.156	.183			.409	.440		.759	.781
		.028	.033		.054	.065	.077	.090			.081	.089		.039	.041



were least efficient when the population relationship was in the middle range, since the conditions of  $\eta^2 = .15$  and  $.40$  had the overall largest standard deviations. Using Formula 14 it can be shown the maximum standard error of  $\epsilon^2$  occurs at about  $\eta^2 = .28$ ,  $\eta^2 = .31$ , and  $\eta^2 = .33$  for  $N = 15$ ,  $N = 30$ , and  $N = 90$  respectively with standard errors of approximately  $.232$ ,  $.151$ , and  $.083$ .

In addition to  $\epsilon^2$  and  $\hat{\omega}^2$ ,  $F$  was computed for each simulated experiment and tested for significance. The frequencies of significant  $F$ 's (not given in this paper) lend support to the previously reported negligible effect of heterogeneity of variance on  $F$  under conditions of equal  $n$  (Norton studies, cited in Lindquist, 1953). Furthermore, these results confirm Box's (1964) derivations of the combined effects of unequal  $n$  and heterogeneity of variance on the percentage of significant  $F$ 's. The number of significant  $F$ 's was underestimated for heterogeneous variances and unequal  $n$  when the largest variance was associated with the largest  $n$ . However, for the reverse case (largest variance associated with smallest  $n$ ), the number of significant  $F$ 's was overestimated. This pattern is directly parallel to that reported for  $\epsilon^2$  and  $\hat{\omega}^2$  in Table 2.

### Discussion

The purpose of this investigation was to study the sampling distributions of  $\hat{\omega}^2$  and  $\epsilon^2$ , measures of strength of relationship, in fixed effects analysis of variance designs. The results appear to promote some caution on the part of investigators in the interpretation of  $\hat{\omega}^2$  and  $\epsilon^2$  with small samples. The standard errors of these two statistics appear to be sizeable for small samples as indicated by Table 2. However, it should be pointed out that the large standard deviations with small samples are typical of other correlational indices.

Hays proposed  $\hat{\omega}^2$  would be useful for two major purposes. One is to see if there is a strong relationship present even though a nonsignificant  $F$  was obtained. The second is to test for a trivial relationship even though a significant  $F$  was obtained due to use of large samples. The results reported show that  $\hat{\omega}^2$  and hence  $\epsilon^2$  have only moderate utility for these purposes. By substitution into equation 10 it can be shown that a nonsignificant  $F$  ( $p > .05$ ) could yield an  $\hat{\omega}^2$  as large as  $.277$  for  $N = 15$ . However sizeable  $\hat{\omega}^2$ 's following nonsignificant  $F$  tests with small  $N$  are quite likely to have come from population associations of zero or close to zero due to the large standard errors with small samples. So even though one's best point estimate of the degree of the population association is close to the obtained  $\hat{\omega}^2$ , the large standard errors for  $\hat{\omega}^2$  with small  $N$  reduces its utility for detecting strong relationships following nonsignificant  $F$  tests.

However, using  $\hat{\omega}^2$  or  $\epsilon^2$  to estimate the degree of association following a significant  $F$  with large  $N$  does have strong merit. This increased utility is due primarily to the reduced standard errors with large  $N$ . The mean  $\hat{\omega}^2$  for  $N = 90$  following  $F$  tests significant at or beyond the  $p < .001$  point did vary across the population  $\eta^2$  values. The utility of computing  $\hat{\omega}^2$  or  $\epsilon^2$  following significant  $F$  tests would of course increase with increases in sample size.

Summarizing some of the other results, it was found that  $\hat{\omega}^2$  and  $\epsilon^2$  were very similar, heterogeneity of variances had negligible effects on the estimates under conditions of equal  $n$ , combinations of heterogeneity of variance and unequal  $n$  yielded especially poor estimates, and use of Formula 6 to estimate  $\hat{\omega}^2$  with unequal  $n$  and homogeneity of variance resulted in reasonable estimates, but did increase the negative bias of the estimates.

One of the purposes of this study was to investigate the effects of violating the assumptions under which  $\hat{\omega}^2$  and  $\epsilon^2$  were defined, namely, homogeneity of variance and proportional representativeness of one's sample sizes. The formulas were not developed for such cases but since they probably will be used in such instances the robustness of these two statistics under violation of these assumptions should be known. We do not mean to imply that the statistic is not appropriate when the conditions under which it was defined have been satisfied. For instance if one were to assume the sample sizes were proportionally representative of the population sizes for the unequal sample size conditions, this would alter the population  $\eta^2$  and a little calculation would show that the  $\hat{\omega}^2$  were indeed better estimates of this population value. However this was not the purpose of our unequal  $n$  conditions as we wanted to investigate the effects on the estimators when the sample sizes were not proportionally representative.

## REFERENCES

- Box, G. E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems, 1. Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics*, 1954, 25, 290-302.
- Glass, G. V and Hakstian, A. R. Measures of association in comparative experiments: Their development and interpretation. *American Educational Research Journal*, 1969, 6, 403-414.
- Hays, W. L. *Statistics for psychologists*. New York: Holt, Rinehart, and Winston, 1963.
- Kelley, T. L. An unbiased correlation ratio measure. *Proceedings of the National Academy of Sciences*, 1935, 21, 554-559.
- Lindquist, E. F. *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin Company, 1953.

- Olkin, I. and Pratt, J. W. Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics*, 1958, 29, 201-211.
- Pearson, K. On the correction necessary for the correlation ratio  $\eta$ . *Biometrika*, 1923, 14, 412-417.
- Vaughan, G. M. and Corballis, M. C. Beyond tests of significance: Estimating strength of effects in selected ANOVA designs. *Psychological Bulletin*, 1969, 72, 204-213.

## OBTAINING PAIRED COMPARISONS DATA FROM MULTIPLE RANK ORDERS USING PARTIALLY BALANCED INCOMPLETE BLOCK DESIGNS<sup>1</sup>

RALPH G. STRATON  
University of Sydney

Complete paired comparisons data was obtained by the method of multiple rank order (MRO) in the context of gathering rank order preferences of grade six students, their parents, and their teachers for instructional objectives. Partially balanced incomplete block designs with two associate classes were used in the MRO instruments instead of the usual balanced incomplete block designs.

The use of partially balanced designs may yield several benefits to a researcher including a reduction in the number of blocks of stimuli to be ranked, a measure of the internal consistency of subject's choices, and a concentration of experimental effort upon comparisons of the most critical stimulus pairs. The benefits and the associated costs of using these designs are discussed in the light of the data obtained in the study.

It is recommended that whenever rank orders or paired comparisons data is called for in a study that serious consideration be given to the use of the MRO method. Furthermore, it is suggested that the overall purposes of a study may best be served by the use of a partially balanced rather than a balanced incomplete block design in the MRO method.

In many educational and psychological research studies rank order data is all that is required of individual subjects. The obtained rank orders may hold intrinsic interest, or, if multiple measurements are available they may be used to determine interval scale values of the stimuli using the law of comparative judgment (Guilford, 1954). In these situations it is usual for either the method of rank order or the

---

<sup>1</sup> The author is indebted to Dr. Jack C. Merwin and Dr. James S. Terwilliger for their helpful comments and suggestions.

method of paired comparisons to be used for data collection. However, the method of multiple rank orders provides an alternative method which is not often used but which has much to recommend it. This paper will focus upon some of the features of the multiple rank orders method and will report on its use in collecting rank order preference data using partially balanced incomplete block designs with two associate classes, instead of the more usual balanced incomplete block designs.

### *Method of Multiple Rank Orders*

The method of multiple rank orders was put forward by Gulliksen and Tucker (1961) as a means of obtaining paired comparisons data with a reduction in experimental labor. Their procedure followed an earlier suggestion by Durbin (1951) for reducing the experimental labor and biasing order effects of the method of rank order. Durbin suggested obtaining rankings within subsets of the stimuli or blocks rather than all together. For example, in Table 1, seven stimuli are presented in seven blocks of three stimuli each.

In a sense, the three methods of rank order, multiple rank orders, and paired comparisons form a continuum of ranking methods. In the rank order method all " $n$ " stimuli are ranked together in a single set or block. Thus, if " $b$ " is the number of blocks and " $k$ " is the number of stimuli within a block we have  $b = 1$  and  $k = n$ . In paired comparisons each stimulus is paired with every other stimulus to form  $b = n(n - 1)/2$  blocks with the  $k = 2$  stimuli being "ranked" within each block. The multiple rank orders method lies between these extremes, having  $1 < b < n(n - 1)/2$  blocks with  $2 < k < n$  stimuli within each block. The rankings given to the stimuli within blocks may allow preferences between the members of each stimulus pair to be deduced, thus yielding paired comparisons data.

TABLE 1  
*Balanced Incomplete Block Design for Seven Stimuli*

Block	Stimuli		
1	1	2	3
2	1	4	5
3	7	1	6
4	4	2	6
5	5	7	2
6	3	4	7
7	6	3	5



### *Allocation of Stimuli to Blocks*

Data collection procedures which use the method of multiple rank orders differ not only in the total number of stimuli ( $n$ ) but also in the number of blocks ( $b$ ), the number of stimuli within each block ( $k$ ), and the conditions governing the allocation of stimuli to blocks. Together these factors make up the experimental design which is followed in allocating stimuli to blocks. It will be apparent that the choice of design is critical in that it will determine what comparisons are made and hence the data yielded from the study and the inferences which may be made.

In his original paper Durbin (1951) suggested that the stimuli be allocated to blocks so as to conform to a balanced incomplete block design (see Cochran and Cox, 1957). In making this suggestion he was guided by two conditions: "(a) Each object should occur an equal number of times in the experiment as a whole. (b) The number of times two particular objects occur together in the same block should be the same for all possible pairs of objects ( $p. 85$ )."

These conditions also led Gulliksen and Tucker (1961) to restrict themselves to balanced incomplete block designs in suggesting the use of multiple rank orders for collecting paired comparisons data. It will be apparent that the arrangement of stimuli in Table 1 conforms to a balanced incomplete block design.

The constraints imposed by these conditions may be an advantage or even a requirement in some studies. However, this is not always the case and other designs may be more advantageous in certain circumstances. Coombs (1964), in recognizing this fact, has remarked that: "... there is nothing sacred about incomplete block designs for collecting data being balanced—and an unbalanced or partially balanced incomplete block design might be considered. . . . The fact that various stimuli and combinations of stimuli would be presented a different number of times (unbalanced) might even become a virtue if more information is needed on some pairs than others, as indeed is usually the case ( $p. 346$ )."

Bock and Jones (1968) and Dykstra (1960) have discussed the analysis of certain partially balanced designs.

### *Attributes of the Multiple Rank Orders Method*

A major advantage of the method of multiple rank orders is the reduction in experimental labor which it affords. Gulliksen and Tucker (1961) have estimated that "For twenty or thirty stimuli the (multiple) rank order design takes, for each subject, only one-half to one-fourth of the time required for the complete paired comparisons

(p. 174)." Furthermore, the complexity of the ranking tasks can be kept low by putting restrictions on  $k$ , the number of stimuli in a block.

Multiple rank orders also allow departures from transitivity to be determined. For paired comparisons data departures from transitivity may be determined for any triad of the stimuli. With multiple rank orders, however, certain triads cannot be circular because the stimuli appear together in the same block, i.e., there is forced transitivity within blocks. Nevertheless, Kendall's (1955) "coefficient of consistence," zeta, may still be used as an index of transitivity in many cases. His formulae will not be appropriate for all multiple rank orders designs, however, due to the constraints which these designs place upon the data. Gulliksen and Tucker (1961) assert that the formulae are always applicable when balanced incomplete block designs are used and Straton (1971) has demonstrated their applicability to certain partially balanced designs.

The channel capacity of the multiple rank orders method, although less than that of paired comparisons, is considerably greater than that of the rank order method for many designs (Gulliksen and Tucker, 1961; Coombs, 1964). According to Coombs "Channel capacity indicates how much information a method might carry and thereby provides a measure of relative power (1964, p. 34)."

### *The Use of Partially Balanced Incomplete Block Designs*

In many studies, where the method of multiple rank orders may be suitable, the use of a balanced incomplete block design will not be appropriate. This may be due to the fact that there is no balanced design available for the number of stimuli specified in the study. However, another difficulty is that the number of blocks in a balanced design is always at least as great as the number of stimuli. Thus, experimental labor may be above tolerable limits. Both of these considerations led to the choice of a partially balanced incomplete block design for use in a study recently completed by the author (Straton, 1971).

### *Nature of the Study*

The study was concerned with the rank order preferences of grade six students, their parents, and their teachers for science instructional objectives. Objectives of two different levels of generality were used. The characteristics of the subjects, the nature of the stimuli, and the design of the study all necessitated that the two data collection methods used should be simple, straightforward, and capable of being

completed quickly. The method of rank order was chosen as one method. For the second method both paired comparisons and multiple rank orders using a balanced incomplete block design were ruled out as pilot study subjects found the methods to be too boring and too time-consuming. Furthermore, no balanced incomplete block design was available for use with fourteen stimuli, the number to be used in the study.

A suitable design was located, however, in the extensive tables of partially balanced incomplete block designs with two associate classes given in Bose, Clatworthy and Shrikhande (1954). These designs violate the second of Durbin's (1951) conditions, i.e., some of the possible pairs of stimuli appear together in the same block more often than other pairs. For 14 objectives there are 91 possible pairs. In the design chosen, seven of these pairs occurred together in the same block three times while all other possible pairs appeared only once (see Table 2). However, each stimulus appeared only three times in the design as a whole, thus conforming to the first of Durbin's (1951) conditions.

The chosen design had several important attributes. First, only seven blocks were required, with six stimuli in a block. A balanced design would have required at least twice as many blocks. Second, the seven replicated stimulus pairs occurred together with four different stimuli in each of three blocks. Thus, each pair was associated with every other stimulus once. Third, each stimulus was included in one and only one replicated pair. Fourth, these replicated stimulus pairs allowed an estimate to be made of the internal consistency of the responses of individual subjects. Many of the designs to be found in Bose, Clatworthy and Shrikhande (1954) have similar attributes.

### *Determining Ranks from Multiple Rank Orders Data*

Sets of rank orders, one per block, were the raw data yielded from the multiple rank orders instruments. The implicit pairwise

TABLE 2  
*Design Used for Multiple Rank Orders Instruments*

Block	Stimuli					
	1	2	3	4	5	6
1	1	8	2	9	4	11
2	2	9	3	10	5	12
3	3	10	4	11	6	13
4	4	11	5	12	7	14
5	5	12	6	13	1	8
6	6	13	7	14	2	9
7	7	14	1	8	3	10

preferences among the stimuli were deduced from these rank orders and recorded in a "Matrix of Votes for Complete Data." In this matrix a one in a cell meant that the column stimulus was preferred over the row stimulus. A zero meant that the row stimulus was preferred. Thus, the sum of the entries in the  $(i, j)$ th and the  $(j, i)$ th cells was equal to the number of times the stimuli appeared together in a block in the whole design. For partially balanced designs this sum is not necessarily unity, whereas for balanced designs it would be.

The "Complete" matrix was converted to a "Matrix of Votes for Paired Comparisons Data" by converting all cells to ones and zeros (see Figure 1). This was done by finding the mean of the  $(i, j)$ th and the  $(j, i)$ th cell entries and scoring one for the cell whose entry was greater than this mean, and zero for the cell whose entry was less than this mean. The final obtained rank order of the stimuli was based upon the column vote totals of the "Paired Comparisons" matrix.

### *Test-Retest Reliability*

Tau coefficients were used as an index of agreement between rank orders obtained from individual subjects. The mean tau coefficient was used as an index of test-retest reliability using a one- to two-week interval. Table 3 gives these mean tau values for the three rater groups and two levels of objectives. The proportion of each group for which  $\tau \geq .50$  is also shown. The probability that random responses would yield a  $\tau \geq .50$  is  $p < .012$  according to the sampling distribution of tau generated by Monte Carlo methods for this study.

### *Internal Consistency Reliability*

The internal consistency reliability of a subject's preference for one of a pair of stimuli may be estimated from those pairs which are presented to the subject more than once. It was assumed that the subject had a "true" preference for one of each pair of stimuli and that this preference did not change during the course of the administration of the instrument. Thus, each time he failed to choose the preferred stimulus he made an error and his response showed unreliability. The more frequently chosen stimulus of a pair was considered to be the truly preferred stimulus. Each choice was scored one or zero (choice score) depending on whether the truly preferred stimulus was chosen on that replication.

An index of internal consistency reliability, gamma ( $\gamma$ ), was defined as one minus the ratio of the observed error variance for a subject to the maximum possible error variance, i.e.:

Stimulus Number	Stimulus Number													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1	0	0	0	0	1	0	0	0	0	0	0	0	1
2	0	1	0	0	0	0	0	0	0	0	0	0	0	0
3	1	0	1	0	1	1	0	0	0	0	0	0	1	1
4	1	1	0	0	1	1	1	1	0	1	0	1	1	1
5	1	1	0	0	0	1	0	0	0	0	0	0	1	1
6	0	1	0	0	0	0	0	0	0	0	0	0	0	1
7	1	1	1	0	1	1	0	1	0	0	0	1	1	1
8	1	1	1	0	1	1	0	0	0	0	0	1	1	1
9	1	1	1	1	1	1	1	1	0	1	1	1	1	1
10	1	1	1	0	1	1	1	1	0	0	1	1	1	1
11	1	1	1	1	1	1	1	1	0	0	0	1	1	1
12	1	1	1	0	1	1	0	0	0	1	0	0	1	1
13	1	1	0	0	0	1	0	0	0	0	0	0	0	1
14	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Votes	10	13	7	2	8	11	4	5	0	3	1	6	9	12
Rank Order	4	1	7	12	6	3	10	9	14	11	13	8	5	2

Figure 1. Matrix of votes for paired comparisons data.



TABLE 3

*Test-Retest Reliability: Mean Tau Values for Rank Order and Multiple Rank Orders Data*

Group	Rank Order			Multiple Rank Orders		
	Tau Mean	S.D.	Proportion $\tau \geq .50$	Tau Mean	S.D.	Proportion $\tau \geq .50$
Level I						
Student	.474	.264	.611	.610	.163	.833
Teacher	.654	.287	.611	.790	.182	.889
Parent	.476	.285	.667	.564	.199	.556
Level II						
Student	.549	.304	.778	.600	.214	.778
Teacher	.639	.267	.833	.628	.264	.833
Parent	.549	.219	.667	.655	.268	.722

Note.— $N = 18$  for each cell.

$$\gamma = 1 - \frac{\text{Error Variance}}{\text{Error Variance (Max.)}}$$

For each replicated stimulus pair the variance of the choice scores was obtained and then these variances were summed across pairs to yield the Error Variance term. The Error Variance (Max.) term was obtained in a similar manner except that the choice scores used were those which would result from the most inconsistent response pattern possible. For three replications this would mean two ones and one zero and for five replications three ones and two zeros.

Gamma values were calculated so as to yield the reliability of a single choice between stimulus pairs for a single subject. The reliability of the total votes for a single stimulus (see Figure 1), upon which the final rank order depends, can be determined by applying the Spearman-Brown formula to the gamma values. This index was called gamma prime ( $\gamma'$ ). In the present study 13 choices contributed to each vote total and so the factor 13 was used in the Spearman-Brown formula to obtain gamma prime values.

Table 4 gives the mean pretest values of gamma prime. The proportion of each group for which  $\gamma' \geq .915$  is also shown. The probability that random responses would yield a gamma prime value of .915 or more is  $p < .014$  according to the sampling distribution generated by Monte Carlo methods for this study.

### *Relationship between Zeta and Gamma*

Both the index of transitivity, Kendall's (1955) zeta, and the index of internal consistency reliability, gamma, may be viewed as indices of internal consistency. In this situation one is forced to consider to what

TABLE 4

*Internal Consistency Reliability: Mean Gamma Prime Values  
for Multiple Rank Orders Pretest Data*

Group	Gamma Prime		Proportion $\gamma' \geq .915$
	Mean	S.D.	
Level I			
Student	.872	.102	.500
Teacher	.966	.048	.889
Parent	.906	.084	.639
Level II			
Student	.935	.051	.806
Teacher	.976	.035	.972
Prent	.943	.058	.861

Note.— $N = 36$  for each cell.

extent these indices are related, and this was estimated using the Pearson product moment correlation coefficient,  $r$  (see Table 5).

In spite of the statistical significance of the correlation coefficients (using  $\alpha = .05$ ), in no case is more than 40% of the variance accounted for. Thus, it seems to be worthwhile to attempt to estimate both types of inconsistency. This is not possible using balanced designs unless the design is replicated, resulting in greater experimental labor.

#### *Agreement Across Methods*

The index of agreement used was the mean of the tau coefficients calculated between rank orders obtained by the two methods, rank order and multiple rank orders, for each subject. These mean taus were calculated for the pretest data and are presented in Table 6 together with the proportion of taus within each subject group for

TABLE 5

*Correlation between Zeta and Gamma Values for  
Multiple Rank Orders Pretest Data*

Group	Pearson $r$	$p$ -value
Level I		
Student	.630	$p < .01$
Teacher	.479	$p < .01$
Parent	.598	$p < .01$
Level II		
Student	.395	$.02 < p < .05$
Teacher	.367	$.02 < p < .05$
Parent	.518	$p < .01$

Note.— $N = 36$  for each cell. The  $p$ -values are those associated with a two-tailed test of the significance of the difference of  $r$  from zero.

which  $\tau \geq .50$ . The probability that random responses would yield a  $\tau \geq .50$  is  $p < .005$  according to the sampling distribution of tau generated by Monte Carlo methods for this study.

Tau values were also calculated across both methods and testing sessions, i.e., allowing two, not just one, main sources of error to enter the situation (see Table 6). There seems to be little doubt that the two methods obtained essentially the same rank orders for most subjects.

### Conclusions

When only rank order data is required from individual subjects the method of multiple rank orders has much to recommend it. It can allow great reductions in time and in experimental labor compared with the method of paired comparisons. It can also reduce the complexity of a subject's task, compared with the method of rank order, particularly when the number of stimuli becomes large. For some types of subjects or for some types of stimuli this might mean as few as 14 stimuli. Since a wide range of possible designs for data collection can be used with the method and since indices of transitivity and internal consistency are available for many of these designs, the method of multiple rank orders would appear to be the preferred method of data collection in many situations.

The empirical data reported here, for three subject types and two types of stimuli, is also favorable to the multiple rank orders method. This method yielded rank orders in close agreement with those obtained by the method of rank order, but with generally greater test-retest reliability. Internal consistency reliability was also generally high. Thus, it is surprising that only a few studies using the method of

TABLE 6  
*Agreement between Methods: Mean Tau Values for Pretest  
Data and Reliability Data*

Group	Pretest Data			Reliability Data		
	Mean	S.D.	Proportion $\tau \geq .50$	Mean	S.D.	Proportion $\tau \geq .50$
Level I						
Student	.581	.249	.860	.502	.298	.611
Teacher	.798	.168	.972	.714	.243	.778
Parent	.598	.193	.667	.540	.219	.694
Level II						
Student	.660	.131	.889	.570	.252	.750
Teacher	.872	.121	1.000	.615	.259	.833
Parent	.591	.230	.722	.568	.251	.722

Note.— $N = 36$  for each cell of the Pretest data and  $N = 18$  for each cell of the Reliability data.

multiple rank orders have been located in the educational and psychological research literature. All except one of these studies used a balanced incomplete block design (e.g., see Borgen and Weiss, 1968; Terwilliger, 1963). Only McKeon (1960, 1961) reports having used a partially balanced incomplete block design with two associate classes as was used in the present study. However, in McKeon's design one of the classes was null whereas in the present study all stimulus pairs appeared together in the same block at least once.

There are, of course, costs involved in the use of the method of multiple rank orders. Channel capacity is less than for the method of paired comparisons and the transitivity of certain triads of the stimuli cannot be determined. The use of a partially balanced or unbalanced design carries the further risk of loss of experimental independence in the replications, but this must be offset against the the possibility of concentrating experimental effort upon comparisons of the most critical stimulus pairs.

It is recommended that whenever rank order or paired comparisons data is called for in a study that serious consideration be given to the use of the method of multiple rank orders. Furthermore, it is suggested that the overall purposes of a study may best be served by the use of a design that is partially balanced or unbalanced, rather than a balanced incomplete block design.

## REFERENCES

- Bock, R. D. and Jones, L. V. *The measurement and prediction of judgment and choice*. San Francisco: Holden-Day, 1968.
- Borgen, F. H. and Weiss, D. J. Application of the method of multiple rank orders to the scaling of environmental characteristics. *Proceedings of the 76th Annual Convention of the American Psychological Association*, 1968, 197-198.
- Bose, R. C., Clatworthy, W. H. and Shrikhande, S. S. *Tables of partially balanced designs with two associate classes*. Raleigh, N. C.: North Carolina Agricultural Experiment Station, Technical Bulletin No. 107, 1954.
- Cochran, W. G. and Cox, G. M. *Experimental designs* (2nd ed.) New York: Wiley, 1957.
- Coombs, C. H. *A theory of data*. New York: Wiley, 1964.
- Durbin, J. Incomplete blocks in ranking experiments. *British Journal of Psychology*, 1951, 4, 85-90.
- Dykstra, O., Jr. Rank analysis of incomplete block designs: A method of paired comparisons employing unequal repetition on pairs. *Biometrics*, 1960, 16, 176-188.
- Guilford, J. P. *Psychometric methods*. New York: McGraw-Hill, 1954.

- Gulliksen, H. and Tucker, L. R. A general procedure for obtaining paired comparisons from multiple rank orders. *Psychometrika*, 1961, 26, 173-183.
- Kendall, M. G. *Rank correlation methods* (2nd ed.) London: Charles Griffin and Co., 1955.
- McKeon, J. J. Some cyclical incomplete paired comparison designs. Chapel Hill: *University of North Carolina Psychometric Laboratory Report No. 24*, 1960.
- McKeon, J. J. Measurement procedures based on comparative judgment. Unpublished doctoral dissertation, Chapel Hill: University of North Carolina, 1961.
- Straton, R. G. An investigation of the nature and measurement of preference data for instructional decision-making. Unpublished doctoral dissertation, Minneapolis: University of Minnesota, 1971.
- Terwilliger, J. S. Dimensions of occupational preference. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1963, 23, 525-542.



## ESTIMATING MOMENTS OF UNIVERSE SCORES AND ASSOCIATED STANDARD ERRORS IN MULTIPLE MATRIX SAMPLING FOR ALL ITEM-SCORING PROCEDURES

TEJ N. PANDEY<sup>1</sup> AND DAVID M. SHOEMAKER

Southwest Regional Laboratory for Educational Research and Development

Described herein are formulas and computational procedures for estimating the mean and second through fourth central moments of universe scores through multiple matrix sampling. Additionally, procedures are given for approximating the standard error associated with each estimate. All procedures are applicable when items are scored either dichotomously or polychotomously.

MULTIPLE matrix sampling is a statistical procedure in which a set of  $K$  items (referred to as the item universe) is subdivided into  $t$  subtests containing  $k$  items each with each subtest administered to  $n$  examinees selected randomly from the population of  $N$  examinees. Although each examinee tested is administered only a portion of the  $K$  items, the results from each subtest may be used to estimate the statistics of the universe scores which would have been obtained by administering all  $K$  items to all  $N$  examinees. The advantages of multiple matrix sampling over traditional testing procedures in the estimation of group performance are numerous, have been cited elsewhere (e.g., Osburn, 1967; Lord and Novick, 1968; Shoemaker, 1972), and need not be enumerated here. Of primary concern is the fact that, with few exceptions, computational formulas available currently in multiple matrix sampling assume that individual test items are scored dichotomously. A restriction such as this is relatively minor in the area of achievement testing for the simple reason that here items are typically scored dichotomously. However, if multiple matrix sampling is to be applicable to other measurement instruments or to other item-

<sup>1</sup> Now at the Office of Program Evaluation and Research, California Department of Education, Sacramento.

scoring procedures, more general equations must be made available. Such has been our goal and the reader will find here computational formulas for estimating moments of universe scores which may be used when items are scored either dichotomously or polychotomously. Estimates of higher central moments of universe scores are required if multiple matrix sampling is used to approximate the entire frequency distribution of universe scores. Additionally, procedures for estimating the standard error of the pooled estimate of the mean universe score and pooled estimate of the second through fourth central moment of universe scores when items are scored polychotomously are given.

### *Estimating Moments of Universe Scores*

The equations given here for estimating the moments of universe scores have been derived within the mathematical framework devised originally by Hooke (1954, 1956a, 1956b) and Lord (1960). Although we report only final results here, a discussion of Hooke's methodology and an expanded version of our analysis is available (Pandey and Shoemaker, 1973).

### *Assumptions and Notation*

We define a population matrix  $X = \|x_{IJ}\|$  for  $I = 1, 2, \dots, N$  examinees and  $J = 1, 2, \dots, K$  items where  $x_{IJ}$  denotes the score obtained by examinee  $I$  on item  $J$ . A matrix sample (bisample) taken from the population matrix is denoted as  $X = \|x_{ij}\|$  for  $i = 1, 2, \dots, n$  examinees and  $j = 1, 2, \dots, k$  items. We assume the  $n$  examinees and  $k$  items in the bisample are a random sample from the  $N$  examinee population and the  $K$ -item universe. We assume that the item-scoring procedure involves a continuous scale and is applied uniformly to each item. Scoring items dichotomously or polychotomously are special cases within our framework. We define an examinee's universe score as the sum of his  $K$  item scores.

### *Parameters Estimated*

We seek unbiased estimators of the mean universe score and the second through fourth central moments of universe scores using the data from one matrix sample. Denoting the universe score for examinee  $I$  as  $x_I$ , we define the mean universe score as

$$\mu_1' = \frac{\sum_{I=1}^N x_I}{N} \quad (1)$$

and the  $r$ th central moment of universe scores as

$$\mu_r = \frac{\sum_{i=1}^N (x_i - \mu_1')^r}{N} \quad (2)$$

### Derived Equations

Equations are given below for estimating the moments when both  $N$  and  $K$  are finite. Because the population is frequently very large, we give additionally equations for estimating the second through fourth central moments when  $N$  is infinite. Following these equations, we describe the procedure used in computing the  $D$ 's,  $A$ 's, and  $F$ 's given in our equations.

#### Mean Universe Scores:

$$\hat{\mu}_1' = \frac{K}{nk} \sum_{i=1}^n \sum_{j=1}^k x_{ij} \quad (3)$$

#### Second Central Moment:

$$\hat{\mu}_2 = \frac{(N-1)}{N} \left\{ D_2 + \frac{D_4}{K} \right\} K^2 \quad (4)$$

$$\hat{\mu}_2 = \left\{ D_2 + \frac{D_4}{K} \right\} K^2 \quad (4a)$$

#### Third Central Moment:

$$\hat{\mu}_3 = \frac{(N-1)(N-2)}{N^2} \left\{ A_4 + \frac{3A_8}{K} + \frac{A_{10}}{K^2} \right\} K^3 \quad (5)$$

$$\hat{\mu}_3 = \left\{ A_4 + \frac{3A_8}{K} + \frac{A_{10}}{K^2} \right\} K^3 \quad (5a)$$

#### Fourth Central Moment:

$$\hat{\mu}_4 = \left[ \frac{3(N-1)^3}{N^3} \left\{ F_4 + \frac{4F_{13}}{K} + \frac{2F_{17}}{K^2} + \frac{2F_{18}}{K} + \frac{4F_{25}}{K^2} + \frac{F_{27}}{K^2} + \frac{F_{20}}{K^3} \right\} + \frac{(N-1)(N^2-3N-3)}{N^3} \left\{ F_8 + \frac{2F_{22}}{K} + \frac{3F_{29}}{K^2} + \frac{4F_{31}}{K^2} + \frac{F_{33}}{K^3} \right\} \right] K^4 \quad (6)$$

$$\hat{\mu}_4 = \left\{ 3F_4 + F_8 + \frac{1}{K} (12F_{13} + 6F_{18} + 2F_{22}) + \frac{1}{K^2} (6F_{17} + 12F_{25} + 3F_{27} + 4F_{31}) + \frac{1}{K^3} (3F_{20} + F_{33}) \right\} K^4 \quad (6a)$$

TABLE 1  
*Sigmas ( $\Sigma$ 's) Associated with Table 2*

$\Sigma$	Arithmetic Computation
1	$(x_{++})^2$
2	$\sum_{i=1}^n x_{i+}^2$
3	$\sum_{j=1}^n x_{+j}^2$
4	$\sum_{i=1}^n \sum_{j=1}^n x_{ij}^2$

### Computational Procedures

The procedure for computing the  $D$ 's,  $A$ 's, and  $F$ 's is described here and it should be noted that doing so is not a casual undertaking on a desk calculator. The reader should anticipate using an electronic computer. Although our tables may seem cumbersome initially, they are in a form which is computerized easily.

Three steps are required in calculating the  $D$ 's,  $A$ 's,  $F$ 's and each will be described in detail for calculating the  $D$ 's. The same procedure is used to compute the  $A$ 's and  $F$ 's. Given the matrix sample  $X = \|x_{ij}\|$ , the following three steps are used to compute any  $D$ :

Step 1: Calculate the sums indicated in Table 1. The plus sign (+) given as a subscript denotes that the subscript replaced by the + is summed over all values.

Step 2: Using the sums calculated from Table 1, calculate the  $d$ -statistics using the coefficients and constants given in Table 2. For example,

$$d_2 = \frac{1}{nk^{(2)}} \{ (1)(\Sigma_2) + (0)(\Sigma_3) + (-1)(\Sigma_4) \}$$

where  $k^{(l)} = k(k-1)(k-2) \dots (k-l+1)$ .

TABLE 2  
*Conversion of Sigmas ( $\Sigma$ 's) to  $d$ -Statistics*

$d$	Multiplicative Constant	1	2	3	4
1	$1/(n^{(3)} k^{(3)})$				1
2	$1/(n k^{(2)})$	1	-1	-1	-1
3	$1/(n^{(2)} k)$		1	0	-1
4	$1/(n k)$			1	1

TABLE 3  
Conversion Table for Computing  $D$ 's from  $d$ 's

$D$	$d$			
	1	2	3	4
1	1			
2	-1	1		
3	-1	0	1	
4	1	-1	-1	1

Step 3: Using the  $d$ -statistics from Table 2, calculate the  $D$ -statistics using the coefficients given in Table 3. For example,

$$D_3 = (-1)(d_1) + (0)(d_2) + (1)(d_3).$$

The same procedure is used to calculate the  $A$ 's and  $F$ 's. The  $A$ 's are calculated using Tables 4, 5, and 6; the  $F$ 's, Tables 7, 8, and 9. The summations in Table 4 and 7 require two additional matrices  $Y$  and  $Z$  where  $Y = \|y_{ij}\| = \|x_{ij}^2\|$  and  $Z = \|z_{ij}\| = \|x_{ij}^3\|$ .

TABLE 4  
Sigmas ( $\Sigma$ 's) Associated with Table 5

$\Sigma$	Arithmetic Computation
1	$(x_{++})^3$
2	$\left(\sum_{i=1}^n x_{i+}^2\right)(x_{++})$
3	$\left(\sum_{j=1}^n x_{+j}^2\right)(x_{++})$
4	$\sum_{i=1}^n x_{i+}^3$
5	$\sum_{j=1}^n x_{+j}^3$
6	$\sum_{i=1}^n \sum_{j=1}^n x_{ij} x_{i+} x_{+j}$
7	$(y_{++})(x_{++})$
8	$\sum_{i=1}^n y_{i+} x_{i+}$
9	$\sum_{j=1}^n y_{+j} x_{+j}$
10	$\sum_{i=1}^n \sum_{j=1}^n x_{ij}^3$



TABLE 5  
Conversion of Sigmas ( $\Sigma$ 's) to  $a$ -Statistics

$n$	Multiplicative Constant	1	2	3	4	5	6	7	8	9	10
1	$1/(n^{12}k^{12})$	1	-3	-3	2	2	6	3	-6	-6	4
2	$1/(n^{12}k^{12})$		1	0	-1	0	-2	-1	3	2	-2
3	$1/(n^{12}k^{12})$			1	0	-1	-2	-1	2	3	-2
4	$1/(n^{12}k^{12})$				1	0	0	0	-3	0	2
5	$1/(n^{12}k^{12})$					1	0	0	0	-3	2
6	$1/(n^{12}k^{12})$						1	0	-1	-1	1
7	$1/(n^{12}k^{12})$							1	-1	-1	1
8	$1/(n^{12}k^{12})$								1	0	-1
9	$1/(n^{12}k^{12})$									1	-1
10	$1/(n^{12}k^{12})$										1

### Calculating Pooled Estimates of Parameters

The estimators defined by equations (3) through (6) produce estimates of parameters using the results of *one* subtest or matrix sample. In multiple matrix sampling there are  $t$  subtests and, because of this,  $t$  estimates of each parameter are obtained. Combining or pooling these  $t$  estimates of each parameter into a single estimate is accomplished by

$$\hat{\rho}_{\text{pooled}} = \frac{\sum_{s=1}^t O_s \hat{\rho}_s}{\sum_{s=1}^t O_s} \quad (7)$$

where  $O_s = n_s k_s$ , the number of observations acquired by subtest  $s$ .

### Calculating Standard Errors of Pooled Estimates

Computing the pooled estimates of each parameter is solving only half the problem. Of equal importance is estimating the standard error

TABLE 6  
Conversion Table for Computing  $A$ 's from  $a$ 's

$A$	1	2	3	4	5	6	7	8	9	10
1	1									
2	-1	1								
3	-1	0	1							
4	2	-3	0	1						
5	2	0	-3	0	1					
6	1	-1	-1	0	0	1				
7	1	-1	-1	0	0	0	1			
8	-2	3	2	-1	0	-2	-1	1		
9	-2	2	3	0	-1	-2	-1	0	1	
10	4	-6	-6	2	2	-6	3	-3	-3	1

TABLE 7  
*Sigmas ( $\Sigma$ 's) Associated with Table 8*

$\Sigma$ Arithmetic Computation	$\Sigma$ Arithmetic Computation
1 $(x_{++})^4$	18 $(y_{++}) \sum_{i=1}^n x_{i+}^3$
2 $(x_{++})^3 \sum_{i=1}^n x_{i+}^3$	19 $(y_{++}) \sum_{j=1}^m x_{+j}^3$
3 $(x_{++})^3 \sum_{j=1}^m x_{+j}^3$	20 $(x_{++}) \sum_{i=1}^n y_{i+} x_{i+}$
4 $\left( \sum_{i=1}^n x_{i+}^2 \right)^2$	21 $(x_{++}) \sum_{j=1}^m y_{+j} x_{+j}$
5 $\left( \sum_{j=1}^m x_{+j}^2 \right)^2$	22 $\sum_{i=1}^n x_{i+}^3 y_{i+}$
6 $(x_{++}) \sum_{i=1}^n x_{i+}^3$	23 $\sum_{j=1}^m x_{+j}^3 y_{+j}$
7 $(x_{++}) \sum_{j=1}^m x_{+j}^3$	24 $\sum_{i=1}^n y_{i+} \left( \sum_{j=1}^m x_{ij} x_{+j} \right)$
8 $\sum_{i=1}^n x_{i+}^4$	25 $\sum_{j=1}^m y_{+j} \left( \sum_{i=1}^n x_{ij} x_{i+} \right)$
9 $\sum_{j=1}^m x_{+j}^4$	26 $\sum_{i=1}^n \sum_{j=1}^m x_{ij}^3 x_{i+} x_{+j}$
10 $(x_{++}) \sum_{j=1}^m \left( \sum_{i=1}^n x_{ij} x_{i+} \right) x_{+j}$	27 $(y_{++})^3$
11 $\left( \sum_{i=1}^n x_{i+}^2 \right) \left( \sum_{j=1}^m x_{+j}^2 \right)$	28 $(x_{++})(z_{++})$
12 $(x_{++})^2 y_{++}$	29 $\sum_{i=1}^n y_{i+}^3$
13 $\sum_{j=1}^m \left( \sum_{i=1}^n x_{ij} x_{i+} \right)^2$	30 $\sum_{j=1}^m y_{+j}^3$
14 $\sum_{i=1}^n \left( \sum_{j=1}^m x_{ij} x_{+j} \right)^2$	31 $\sum_{i=1}^n x_{i+} z_{i+}$
15 $\sum_{i=1}^n \left( \sum_{j=1}^m x_{ij} x_{i+} \right) \left( \sum_{j=1}^m x_{ij} x_{+j} \right)$	32 $\sum_{j=1}^m x_{+j} z_{+j}$
16 $\sum_{j=1}^m \left( \sum_{i=1}^n x_{ij} x_{i+} \right) \left( \sum_{i=1}^n x_{ij} x_{+j} \right)$	33 $\sum_{i=1}^n \sum_{j=1}^m x_{ij}^4$
17 $\sum_{j=1}^m \left( \sum_{i=1}^n x_{ij}^2 \right)^2 + \sum_{j=1}^m \sum_{h=1}^m \left( \sum_{i=1}^n x_{ij} x_{ih} \right)^2$	

of estimate associated with this value. Two general procedures are available for estimating the standard error of estimate. The first procedure is one described originally by Lord and Novick (1968) and uses results reported by Hooke (1954, 1956a, 1956b); the second, referred to as the "jackknife" procedure, has been popularized by



TABLE 8 (Continued)

$\Sigma$	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
-24	-24	4	6	-6	-6	-24	-24	36	36	24	24	48	3	16	-18	-18	-48	-48	36
6	6	4	-2	2	1	6	4	-12	-12	-6	-6	-12	-1	-4	6	6	16	12	-12
4	4	6	-2	1	2	4	6	-6	-12	-8	-8	-12	-1	-4	6	6	12	16	-12
0	0	0	2	-2	0	0	0	6	0	0	8	0	1	0	-3	-6	-8	0	6
0	0	0	2	0	-2	0	0	6	6	8	0	0	1	0	-6	-3	0	-8	6
-3	0	-3	0	0	0	-3	-3	0	6	3	0	6	0	2	-3	0	-8	-6	6
0	0	0	0	0	0	0	0	-6	0	0	0	0	0	0	3	0	-6	-8	6
0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	3	8	0	6
-1	-1	-1	1	0	0	-1	-1	2	2	2	2	3	0	1	0	-2	-4	-4	4
-2	-2	0	0	-1	-1	0	0	2	2	2	2	4	1	0	-2	-2	-4	-4	4
0	0	0	0	0	0	-2	-2	2	2	2	2	2	1	2	-2	-2	-4	-4	4
0	0	0	-1	0	0	0	0	1	0	0	-2	0	0	0	1	2	2	0	-2
0	0	0	-1	0	0	0	0	0	-1	-2	0	0	0	0	2	1	0	2	-2
1	1	0	0	0	0	0	0	-1	0	-1	0	-2	0	0	1	0	2	2	-2
			0	0	0	0	0	0	-1	0	-1	-2	0	0	0	1	2	2	-2
			1	0	0	0	0	0	0	0	0	0	0	0	-1	-1	0	0	1

$f$	Multiplicative	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
18	Constant																	
19	$1/(n^{(1)}k^{(1)})$																	
20	$1/(n^{(2)}k^{(2)})$																	
21	$1/(n^{(3)}k^{(3)})$																	
22	$1/(n^{(4)}k^{(4)})$																	
23	$1/(n^{(5)}k^{(5)})$																	
24	$1/(n^{(6)}k^{(6)})$																	
25	$1/(n^{(7)}k^{(7)})$																	
26	$1/(n^{(8)}k^{(8)})$																	
27	$1/(n^{(9)}k^{(9)})$																	
28	$1/(n^{(10)}k^{(10)})$																	
29	$1/(n^{(11)}k^{(11)})$																	
30	$1/(n^{(12)}k^{(12)})$																	
31	$1/(n^{(13)}k^{(13)})$																	
32	$1/(n^{(14)}k^{(14)})$																	
33	$1/(n^{(15)}k^{(15)})$																	

18	1	0	0	0	0	0	-1	0	0	-2	0	-1	0	0	1	1	2	0	-2	33
19	1	1	0	0	0	0	0	-1	-2	0	0	-1	0	0	0	0	0	2	-2	32
20			1	0	0	0	-1	0	-1	0	-1	0	0	-1	0	0	2	2	-2	31
21				1	0	0	0	-1	0	-1	0	0	0	-1	0	0	2	2	-2	30
22					1	0	0	0	0	0	0	0	0	0	0	0	0	2	2	29
23						1	0	0	0	0	0	0	0	0	0	0	0	0	2	28
24							1	0	0	0	0	0	0	0	0	0	0	0	0	27
25								1	0	0	0	0	0	0	0	0	0	0	0	26
26									1	0	0	0	0	0	0	0	0	0	0	25
27										1	0	0	0	0	0	0	0	0	0	24
28											1	0	0	0	0	0	0	0	0	23
29												1	0	0	0	0	0	0	0	22
30													1	0	0	0	0	0	0	21
31														1	0	0	0	0	0	20
32															1	0	0	0	0	19
33																1	0	0	0	18



Mosteller and Tukey (1968). Although both procedures may be used when items are scored polychotomously, the first is applicable only for those item-sampling plans having the product  $tk$  less than or equal to  $K$ . The jackknife procedure may be used with *all* item-sampling plans.

### *The Hooke-Lord-Novick Equations*

If items and examinees have been sampled randomly, the squared standard error of estimate of the mean universe score for subtest  $s$  is equal to

$$\text{VAR } (\hat{\mu}_{1s}') = \left[ \frac{D_1}{n_s k_s} + \frac{D_2}{n_s} + \frac{(K - k_s) D_3}{K k_s} \right] K^2. \quad (8)$$

Equation (8) estimates the error variance associated with the estimate of the mean universe score obtained from one subtest. The estimate of the standard error of the *pooled* estimate of the mean, however, is a function of equation (8) and is computed as

$$\text{SE } (\hat{\mu}_{1_{\text{pooled}}}') = \frac{1}{t} \left\{ \sum_{s=1}^t \text{VAR } (\hat{\mu}_{1s}') - \frac{(t-1) K^2 \sum_{s=1}^t \hat{\sigma}_{ps}^2}{(K-1)} \right\}^{1/2} \quad (9)$$

where  $\hat{\sigma}_{ps}^2$  refers to the estimate of the variance of the mean item scores (variance of the item difficulty indices) for subtest  $s$ .

The squared standard error of estimate for the variance of universe scores for subtest  $s$  is equal to

$$\begin{aligned} \text{VAR } (\hat{\mu}_{2s}) = & \left[ \frac{2F_4}{n_s - 1} + \frac{F_8}{n_s} + \frac{4n_s F_{13}}{k_s(n_s - 1)} + \frac{2n_s F_{17}}{k_s(k_s - 1)(n_s - 1)} \right. \\ & + \frac{4F_{18}}{k_s(n_s - 1)} + \frac{4F_{23}}{n_s k_s} + \frac{2F_{27}}{k_s(k_s - 1)(n_s - 1)} \\ & \left. + \frac{2F_{29}}{n_s k_s(k_s - 1)} \right] K^4 \end{aligned} \quad (10)$$

The standard error of the pooled estimate of the variance is computed as

$$\text{SE } (\hat{\mu}_{2_{\text{pooled}}}) = \frac{1}{t} \left\{ \sum_{s=1}^t \text{VAR } (\hat{\mu}_{2s}) - (t-1) 4K \sum_{s=1}^t \text{VAR } (s_{kz})_s \right\}^{1/2} \quad (11)$$

where  $\text{VAR } (s_{kz})_s$  refers to the variance of the covariances between each item and the mean item score for subtest  $s$ . It should be noted that equations (9) and (11) were derived under the assumption that the number of examinees in the population is infinitely large.

Equations comparable to (8) and (10) could be derived for the third and fourth central moments within the framework developed by

TABLE 9  
Conversion Table for Computing  $F_s$  from  $f_s$

$F$	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1													
2	-1	1												
3	-1	0	1											
4	1	-2	0	1										
5	1	-3	0	0	1									
6	2	0	-3	0	0	1								
7	2	12	0	-3	0	-4	1							
8	-6	0	12	0	-3	0	0	1						
9	-6	-1	-1	0	0	0	0	0	1					
10	1	-1	-1	0	0	0	0	0	0	1				
11	1	-1	-1	0	0	0	0	0	0	0	1			
12	1	-1	-1	0	0	0	0	0	0	0	0	1		
13	-1	2	1	-1	0	0	0	0	0	0	0	0	1	
14	-1	1	2	0	-1	0	0	0	0	0	0	0	0	1
15	-2	3	2	0	0	0	0	0	0	0	0	0	0	0
16	-2	2	3	0	0	-1	0	0	0	0	0	0	0	0
17	1	-2	-2	1	0	0	-1	0	0	0	0	0	0	0
18	-1	2	1	-1	0	0	0	0	0	0	0	0	0	0
19	-1	1	2	0	-1	0	0	0	0	0	0	0	0	0
20	-2	3	2	0	0	-1	0	0	0	0	0	0	0	0
21	-2	2	3	0	0	0	-1	0	0	0	0	0	0	0
22	6	-12	-6	3	0	0	0	-1	0	0	0	0	0	0
23	6	-6	-12	0	3	0	0	0	-1	0	0	0	0	0
24	2	-3	-4	0	2	1	0	0	0	0	0	0	0	0
25	2	-4	-3	2	0	0	1	0	0	0	0	0	0	0
26	4	-6	-6	0	1	2	2	0	0	0	0	0	0	0
27	1	-2	-2	1	0	0	0	0	0	0	0	0	0	0
28	4	-6	-6	0	1	2	2	0	0	0	0	0	0	0
29	-6	12	12	-3	-6	-4	0	1	0	0	0	0	0	0
30	-6	12	12	-6	-3	0	0	0	0	0	0	0	0	0
31	-12	24	18	-6	0	0	-4	0	1	-16	-4	-4	4	8
32	-12	18	24	0	-6	-6	-8	2	0	-24	-6	-6	6	6
33	36	-72	-72	18	-18	-24	24	-6	-6	96	24	-6	-24	-24

TABLE 9 (Continued)

[illegible]

Hooke; however, doing this would not be a casual undertaking. In Hooke's procedure, estimating the standard error of the  $i$ th moment requires terms related to the  $2i$ th moment and the number of sums, i.e.,  $\Sigma$ s in Tables 1, 4, and 7, becomes large. For example, for the 5th, 6th, 7th and 8th moments, 91, 298, 910, and 3017 sums, respectively, must be computed. Were one to pursue this line of research, the work of Dayhoff (1964, 1966) will be of use.

### *The Jackknife Procedure*

The jackknife procedure provides an alternative procedure for computing standard errors of pooled estimates in multiple matrix sampling. The jackknife procedure could be used to compute the standard errors of the pooled estimates of the third and fourth central moments after the results from each subtest had been collected. The computations involved in the jackknife are relatively simple. Let

$y_{all}$  = the pooled estimate of the parameter using all subtests, and  
 $y_{(j)}$  = the pooled estimate of the parameter computed after removing subtest  $j$ .

Defining

$$y^*_j = ty_{all} - (t-1)y_{(j)} \quad \text{for } j = 1, 2, \dots, t$$

the jackknifed estimate of the parameter is equal to

$$y^* = (y^*_1 + y^*_2 + \dots + y^*_t)/t \quad (12)$$

with an estimate of its variance given by

$$\text{VAR}(\hat{\rho}_{\text{pooled}}) = \frac{\sum_{j=1}^t (y^*_j - y^*)^2}{t(t-1)} \quad (13)$$

Shoemaker (1973) has verified empirically that the jackknife procedure may be used to approximate standard errors of estimate in multiple matrix sampling. It should be noted, however, that when the variance of the item difficulty indices is greater than zero, the jackknife procedure estimates conservatively the standard error of the *mean-universe score*.

### *Conclusion*

If multiple matrix sampling is to be used more widely, computational formulas which incorporate all uniform item-scoring procedures must be available and in a form easy to compute. Our results are a step in this direction. Although at first glance the tables

may seem cumbersome to use, they are in a form which lends itself readily to being programmed on a computer.

## REFERENCES

- Dayhoff, E. On the equivalence of polykays of the second degree and sigmas. *Annals of Mathematical Statistics*, 1964, 35, 1663-1672.
- Dayhoff, E. Generalized polykays, an extension of simple polykays and bipolykays. *Annals of Mathematical Statistics*, 1966, 37, 226-241.
- Hooke, R. *The estimation of polykays in the analysis of variance*. Princeton University: Statistical Research Group, Memorandum No. 56, 1954.
- Hooke, R. Symmetric functions of a two-way array. *Annals of Mathematical Statistics*, 1956, 27, 55-79. (a)
- Hooke, R. Some applications of bipolykays to the estimation of variance components and their moments. *Annals of Mathematical Statistics*, 1956, 27, 80-98. (b)
- Lord, F. M. Use of true-score theory to predict moments of univariate and bivariate observed-score distributions. *Psychometrika*, 1960, 25, 325-342.
- Lord, F. M. and Novick, M. R. *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968.
- Mosteller, F. and Tukey, J. W. Data analysis, including statistics. In Lindzey, G. and Aronson, E. (Eds.) *The handbook of social psychology*. (Second ed.) Reading, Mass.: Addison-Wesley, 1968.
- Osburn, H. G. A note on design of test experiments. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1967, 27, 297-302.
- Pandey, T. N. and Shoemaker, D. M. *Estimating moments of universe scores and associated standard errors in multiple matrix sampling for all item-scoring procedures*. Southwest Regional Laboratory for Educational Research and Development, Technical Memorandum, 1973.
- Shoemaker, D. M. Evaluating the effectiveness of competing instructional programs. *Educational Researcher*, 1972, 1, 5-8.
- Shoemaker, D. M. A note on allocating items to subtests in multiple matrix sampling and approximating standard errors of estimate with the jackknife. *Journal of Educational Measurement*, 1973, 10, 311-319.





## THE STRUCTURE OF DOMAIN HIERARCHIES FOUND WITHIN A DOMAIN REFERENCED TESTING SYSTEM<sup>1</sup>

GEORGE B. MACREADY  
University of Maryland

The purpose of this study was to assess conditional states of item mastery found among items from different item domains and the effectiveness of various procedures for identifying such conditional relations. The item domains considered were from the curriculum area of multiplication of whole numbers, and were defined by a domain referenced testing system. It was possible to infer from the results of this study that the domain referenced testing system considered produced items which across domains showed strong conditional relations. Comparisons of goodness of fit were made among domain hierarchies with similar numbers of specified conditional relations generated by two different empirical procedures and by experts judgment. Additional comparisons were made among models generated by the same procedure but with different numbers of specified conditional relations.

Support for the validity of empirically generated hierarchies with moderate numbers of conditional relations among domains was provided. However, similar support was not provided for the conditional relations hypothesized to exist by subject matter experts.

ONE kind of information that may prove to be useful to the educator in making decisions about the implementation and improvement of an instructional design deals with the nature of the underlying structure of the subject matter (e.g., Resnick and Wang, 1969). This structure deals with the order of acquisition found among specified intellectual

---

<sup>1</sup> This study was partially supported by grants from the General Research Board and the Bureau of Educational Research and Field Services, University of Maryland. The author is grateful to Jack C. Merwin for helpful comments on an earlier draft of this paper.

capabilities and is described as a "hierarchy" among these skills (Gagne, 1962; Glaser and Nitko, 1971). The specifications for such hierarchies may vary from one extreme called a linear ordered set of variables, which imposes maximum restrictions regarding order, to the other extreme called a completely independent set of variables, in which no restrictions are imposed on the order of the variables.

In education today practitioners either explicitly or implicitly make assumptions about the underlying structure of the subject matter. However, they seldom go on to empirically verify or modify their conceptions of the subject matter. Instead, much of the strengthening or modifying of their original conceptions are based on an intuitively guided process. Such procedures may lead to rigid, overly simplified, and stilted conceptions of the subject matter.

Empirical study in this area is difficult, primarily since the subject-matter areas are not seen as being readily assessable to such analysis because of a lack of clarity regarding content. One means of dealing with the lack of clarity of content found in many educational achievement tests is to use a domain referenced approach to testing. Following Hively's (1970) conceptualization, the universe of items which is to be considered is specified in operational terms by means of rules called *item form rules*. Specific sets of these rules are used to specify within which of a number of possible subsets each item contained in the universe falls. The subsets of items are called *domains*.

In the specification of the item form rules an attempt is made to place items within a given domain in such a way that each item within the domain is testing the same underlying skills. It is then possible to describe a set of logical relations among the various domains of items. Such a set of relations is called a *domain hierarchy*.

### *Method*

#### *Instrument Construction, Administration and Scoring*

The domain referenced test used in this study was the section of Honeywell's "Arithmetic Test Generation Program" (ATG), dealing with multiplication of whole numbers (see Patterson and Vierling, 1970).

The ATG program has grouped multiplication items into 20 non-overlapping domains. The six domains used in the study were chosen on the basis of pilot study results.

In the pilot study, a multiplication test was administered to a total of 115 fifth grade students. The test was composed of one randomly generated item from each of the 20 domains provided by the ATG

Program. To obtain a more representative sample of the items within each domain, three different forms of the pilot test composed of different randomly generated items were administered to one-third of the students. In the selection of the six domains that were considered in the main study, operationally defined selection rules were used such that the domains selected had average item difficulties uniformly distributed between .3 and .7. This was done to obtain a large amount of subject variability on items within the domains (see Macready and Merwin, 1973) and to provide an opportunity to identify a wide assortment of hierarchies which might exist among domains. A description of the characteristic skills involved in working items from the six selected domains are listed in Table 1.

For use in the main study, ten items were randomly sampled from each of the six selected domains. The 60 items thus generated made up the specific content of the test used. This test was administered to 285 students in 10 fifth-grade classrooms in the Minneapolis public schools. The items on the test were then dichotomously scored as either right or wrong.

### *Generation of Hierarchical Structures*

The generation of hypothetical hierarchies among the domains considered in this study were carried out in an attempt to reflect ordered relations in the acquisition of skills necessary to work the items from the various domains. To generate such hierarchies, both "theoretical" and "empirical" procedures were used.

First, a theoretically generated hierarchy was considered. This

TABLE 1  
*Item Form Rules and Characteristic Items Found in the Domains Studied<sup>a</sup>*

Domain No.	Prototype Item	Item Format	Characteristic Skill
10	824	A	2 digit multiplier; no carry
12	$\begin{array}{r} \times 21 \\ \hline 432 \end{array}$	$\begin{array}{r} \times B \\ \hline A \end{array}$	2 digit multiplier; multiple of 10
13	$\begin{array}{r} \times 40 \\ \hline 347 \end{array}$	$\begin{array}{r} \times B \\ \hline A \end{array}$	2 digit multiplier; easy carry
15	$\begin{array}{r} \times 23 \\ \hline 8647 \end{array}$	$\begin{array}{r} \times B \\ \hline A \end{array}$	2 digit multiplier; hard carry
17	$\begin{array}{r} \times 69 \\ \hline 627 \end{array}$	$\begin{array}{r} \times B \\ \hline A \end{array}$	3 digit multiplier with middle digit equal to 0.
18	$\begin{array}{r} \times 204 \\ \hline 472 \end{array}$	$\begin{array}{r} \times B \\ \hline A \end{array}$	3 digit multiplier with no digits equal to 0.
	$\begin{array}{r} \times 361 \\ \hline \end{array}$	$\begin{array}{r} \times B \\ \hline \end{array}$	

<sup>a</sup>The item form rules used to define the items found within each domain are presented by Macready (1972).

hierarchy was based on mathematics teachers' professional judgment as to what preliminary skills are necessary prerequisites for the learning of more advanced skills. The specific hierarchy considered was from the ATG manual (see Patterson and Vierling, 1970) and involved the 20-item domains found in the area of "multiplication of whole numbers." For the purposes of this study, only that portion of the hierarchy which dealt with the domains used was considered. This portion of the hierarchy is schematically represented in Figure 1. Such a schematic representation of a hierarchy may be interpreted in the following manner. The numbers to the right of each small circle are the identification numbers for the domains which the small circles represent. The lines connecting the small circles represent conditional relations that hold between the domains whose representative circles are connected. The domain represented by the lower of the two connected circles is considered to be a necessary prerequisite to the domain represented by the higher circle. Thus, the acquisition of skills necessary for correctly working the items from a lower level domain is considered necessary but not sufficient for the acquisition of the skills necessary for correctly working the items from a connected higher level domain. Another important characteristic of these schematic structures is that the conditional relations which are represented are transitive. Schematically, this means that any two circles which are indirectly connected by a continuously increasing or decreasing set of line segments are considered to have a conditional relation existing between their domains, such that the higher of the two is considered conditional on the acquisition of the lower. In the hierarchy represented above this means that both domains 10 and 13 are considered to be conditional prerequisites for the acquisition of domains 12, 15, and 18, while for domain 17, the conditional prerequisites are considered to be domains 10, 12, and 13.

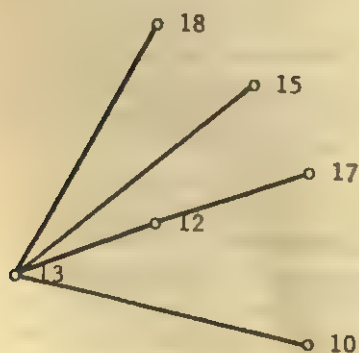


Figure 1. Theoretically generated hierarchy based on mathematicians' professional judgment.



In addition to the theoretical hierarchy presented above, a number of empirically generated hierarchies based on the pilot study data were also considered. There were two methods used in generating these hierarchies.

The first method was a slight modification of a procedure suggested by Carroll and described by Resnick and Wang (1969). The procedure determines the ordered relations to be considered on the basis of item indices of homogeneity,  $H_{ij}$ , found between items. The statistic,  $H_{ij}$ , estimates the proportion of possible increase in the probability of getting item  $i$  correct, given that a more difficult item  $j$  is answered correctly. This statistic is equivalent to  $\Phi_i/\Phi_{\max}$ , which was used by Carroll.

Since in the case of the pilot data only one item from each domain was administered to each student. The procedure used consisted of identifying the  $H_{ij}$  indices for all pairs of items and then specifying the conditional relations which were to be considered in the hierarchy on the basis of the magnitude of these indices. This was done by taking the  $K$  largest coefficients of  $H_{ij}$  and specifying conditional relations between the domains involved, such that the more difficult domain in the pair was considered conditional on the less difficult domain. One exception to the above procedure was that a conditional relation was not considered unless it showed transitivity with the other conditional relations being considered. This situation occurred only once. This was in the case of the " $H_{ij}$ " generated model based on 10 specified conditional relations. To obtain transitivity in this model, the nontransitive  $H_{ij}$  value was replaced (i.e., the 10 largest transitive  $H_{ij}$  values were used for generating the model).

By using the above procedure and varying the minimum acceptable value of  $H_{ij}$  for inclusion of the conditional relations, six different hierarchies were generated. The cutting points for acceptable magnitudes of  $H_{ij}$  were chosen in such a way as to allow for comparisons with the hierarchies with similar numbers of conditional restrictions generated by the other procedures and to allow for comparisons among various structures generated by this procedure with varying numbers of conditional relations between the domains. These hierarchies are schematically represented in Figures 2 through 7.

The second general method used to empirically generate domain hierarchies from the pilot data was based on a method suggested by Bart and Krus (1973). In this procedure, the response patterns of individual students were used in determining what particular hierarchical relations among domains seems to best fit the data. This was done by considering the item response patterns for each individual on the item from each domain in some specified order. Those patterns

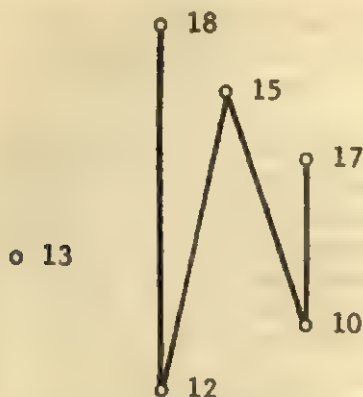


Figure 2. Generated hierarchy based on the four largest  $H_{ij}$  values. (Note: The minimum value of  $H_{ij}$  between domains with specified conditional relations is equal to .73.)

which some specified minimum proportion of subjects had obtained were selected and used to generate a domain hierarchy. The hierarchies generated had the minimum possible number of specified conditional relations which were still "compatible" with the response patterns being considered. For a hierarchy to be compatible with a set of response patterns it is necessary that the conditional relations specified by the hierarchy allow all of the selected patterns of response to occur without the necessity of using an error component. By using this procedure and varying the minimum proportion of subjects necessary for the selection of a response pattern, three different hierarchies were generated. It should be noted that all possible criteria of minimally acceptable proportions of subjects for inclusion of response patterns were considered. However, criteria of minimum proportion of  $S$ s obtaining a given response pattern falling below .02 were not used. This was because under such criteria the acceptable response patterns did

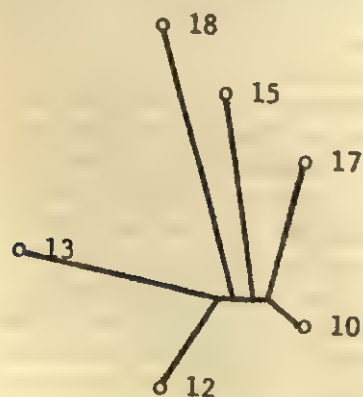
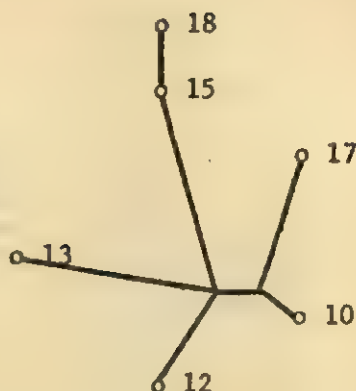


Figure 3. Generated hierarchy based on the eight largest  $H_{ij}$  values. (Note: The minimum value of  $H_{ij}$  between domains with specified conditional relations is equal to .63.)

Figure 4. Generated hierarchy based on the nine largest  $H_{ij}$  values. (Note: The minimum value of  $H_{ij}$  between domains with specified conditional relations is equal to .58.)



not provide for any acceptable conditional relations between domains. Criteria of minimum proportion of greater than .04 were also rejected since the number of response patterns falling above .04 was very small. The hierarchies generated are schematically represented in Figures 8 through 10.

### Results

#### *Relations among Items within Domains*

The mean item scores,  $\bar{p}$ , found within each domain showed considerable spread in magnitude. The specific values of  $\bar{p}$  for each domain were:  $\bar{p}(10) = .71$ ,  $\bar{p}(12) = .70$ ,  $\bar{p}(13) = .61$ ,  $\bar{p}(15) = .43$ ,  $\bar{p}(17) = .53$  and  $\bar{p}(18) = .41$ . However, the standard deviation of item difficulties found within the various domains were relatively small, (rang-

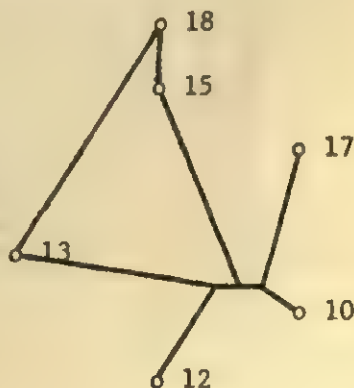


Figure 5. Generated hierarchy based on the ten largest transitive  $H_{ij}$  values. (Note: The minimum value of  $H_{ij}$  between domains with specified conditional relations is equal to .54.)

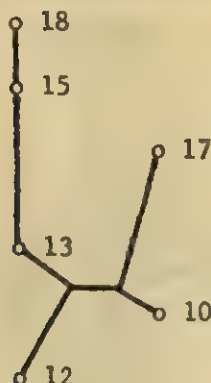


Figure 6. Generated hierarchy based on the eleven largest  $H_{ij}$  values. (Note: The minimum value of  $H_{ij}$  between domains with specified conditional relations is equal to .54.)

ing from .02 to .09), when compared to the same statistic across domains, which was .13. At the same time coefficients of internal consistency (i.e.,  $KR-20$ ) among items within the various domains were found to be high. These coefficients ranged from .84 to .93.

#### *Analysis of the Domain Hierarchy Models*

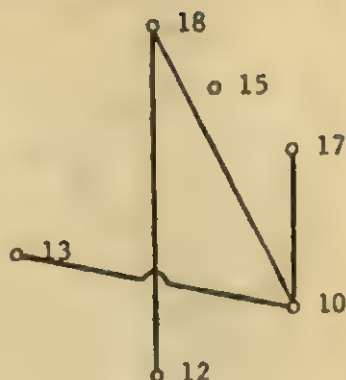
Two general kinds of comparisons of "goodness of fit" provided by the various generated hierarchical models to the actual data were considered. First, comparisons were made among hierarchical models within each hierarchical generation procedure with varying numbers of conditional relations. Second, comparisons were made among the models with similar numbers of specified conditional relations produced by the various generation procedures.

The kind of evidence which was used as a means of comparing the



Figure 7. Generated hierarchy based on all of the  $H_{ij}$  values. (Note: The minimum value of  $H_{ij}$  between domains is equal to .45.)

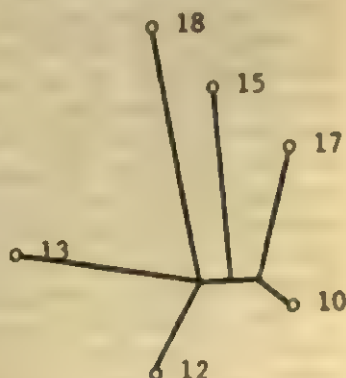
Figure 8. Generated hierarchy based on the seventeen most frequently occurring response patterns (Note: The minimum proportion of *Ss* obtaining a given response pattern for inclusion of that pattern was set at .02.)



various generated models was mean item indexes of homogeneity,  $H_{ij}$ . These values were based on all pairs of items which were from different domains having hypothesized conditional relations between them (the  $H_{ij}$  values which were used in this section were slightly modified in that the hypothesized conditional relations were used in determining the direction of the conditional probability within  $H_{ij}$  rather than the relative difficulties of the items).

The results in Table 2 show that the  $H_{ij}$  values based on the various generated models, (see columns designated I), provide quite substantial coefficients when compared with the mean of all possible  $H_{ij}$  values, .415 (these values were based on both of the possible conditional relations between all pairs of items from different domains), or when compared to the mean of  $H_{ij}$  values based on conditional relations not suggested by the models, (see columns designated III). It may further be noted that, within each of the hierarchical generation

Figure 9. Generated hierarchy based on the twelve most frequently occurring response patterns. (Note: The minimum proportion of *Ss* obtaining a given response pattern for inclusion of that pattern was set at .03.)





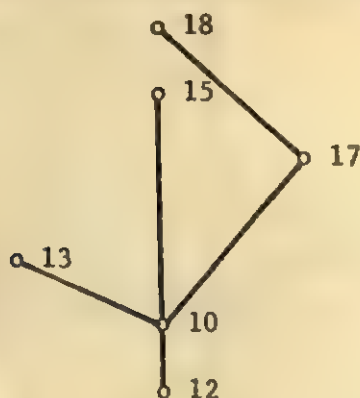


Figure 10. Generated hierarchy based on the seven most frequently occurring response patterns. (Note: The minimum proportion of *Ss* obtaining a given response pattern for inclusion of that pattern was set at .04.)

procedures, as the number of conditional relations between domains was decreased there was an increase in the  $\bar{H}_{ij}$  values suggested by the corresponding model. However, one may also note that there was a simultaneous increase in the  $\bar{H}_{ij}$  values based on all other conditional relations not suggested by the model.

It next became of interest to determine the number of conditional relations which could effectively be specified within a given generation procedure. This was accomplished by assessing the "strength" of conditional relations found within a given model but not found in similarly generated models with fewer specified relations (i.e., the strength of conditional relations which were added to previously specified relations were assessed). To determine the strength of the added conditional relations,  $\bar{H}_{ij}$  values based solely on the added relations were used. These values are also presented in Table 2 (see columns designated II).

When this approach was used for the " $H_{ij}$ " generated models, it was found that the added conditional relations provided relatively large  $\bar{H}_{ij}$  values for those models with eight or fewer specified relations, while those models with more than eight specified relations provided corresponding values which were markedly smaller. Similar results were found in the case of the "response frequency" generated models. However, in this case, the model with 10 specified relations showed a much less dramatic drop than that found in the corresponding " $H_{ij}$ " generated model (the  $\bar{H}_{ij}$  values for the two added conditional relations were .44 and .53, respectively, for the " $H_{ij}$ " and the "response frequency" generated models with 10 specified relations).

It was of further interest to note the differences in magnitude found between  $\bar{H}_{ij}$  values based on added conditional relations within the models and the corresponding  $\bar{H}_{ij}$  values based on conditional rela-

TABLE 2  
Comparisons of  $H_{ij}$  Values Based on Conditional Relations Specified by the Various Generated Models

No. of Conditional Relations	Models											
	" $H_{ij}$ " generated $H_{ij}$				"Response Frequency" generated $H_{ij}$				"Theoretically" generated $H_{ij}$			
	I	II	III	IV	I	II	III	IV	I	II	III	IV
30 <sup>a</sup>	.41	—	—	—	—	—	—	—	—	—	—	—
15	.54	.48	.29	.19	—	—	—	—	—	—	—	—
11	.56	.48	.33	.15	—	—	—	—	—	—	—	—
10	.57	.52	.34	.14	.59	.53	.33	.20	.54	.54	.35	.19
9	.58	.35	.35	.00	—	—	—	—	—	—	—	—
8	.61	.61	.34	.27	.61	.61	.34	.27	—	—	—	—
4	.61	.61	.39	.22	.61	.61	.39	.22	—	—	—	—
0 <sup>a</sup>	—	—	.41	—	—	—	—	—	—	—	—	—

Note.—Where: I = The mean of the  $H_{ij}$  values based on the conditional relations specified by the particular model in question. II = The mean of the  $H_{ij}$  values based on conditional relations specified within the particular model but not found within similarly generated models with fewer specified conditional relations. III = The mean of the  $H_{ij}$  values based on the conditional relations not specified by the particular model in question. IV = The difference between II and III, above

<sup>a</sup> These are the limiting cases for number of specified conditional relations.

tions outside of the models. The differences between these values are presented in Table 2, (see columns designated IV). Comparisons of these difference values showed that the model with eight or less specified conditional relations provided much larger differences than the other models. This was true for both of the empirical generation procedures.

Comparisons were also made among models generated by different generation procedures with equivalent numbers of specified relations. This was done by comparing the  $\bar{H}_{ij}$  values based on all of the specified relations within the models. Differences were only found in the case of those models with 10 specified relations. Here the "response frequency" generated model provided the largest value followed by the " $H_{ij}$ " generation model. The "theoretical" generated model was a low third, providing the lowest  $\bar{H}_{ij}$  value of any of the models generated, including those models with larger numbers of specified relations.

### *Discussion*

The results of this investigation suggested that the specified conditional relations, for all of the models considered, provided relatively good "fit" to the data (this lends support to the contention that students tend to learn how to correctly work items within specified domains prior to items in other domains). This assessment was inferred from the fact that the proportions of possible decreases in item difficulties based on the conditional information specified by the models were considerably larger than similar proportions based on non-specified conditional information (i.e., the effect of "success" on items from a given domain more greatly affected the probability of "success" on items from domains specified as being conditional on "success" in the first domain).

This finding of relatively "good fit" across all hierarchies considered is not surprising since all of the generated models specified sets of conditional relations which were quite similar to one another. These similarities among specified conditional relations were found both with respect to the specific domains involved and with respect to the direction of the conditional relations. One of the most noticeable characteristics, which was found to play a prominent role within all of the models generated, dealt with domains 10 and 12. This was because all of the generated models tended to list frequently these two domains as prerequisites for the acquisition of the other domains. Another characteristic which was found throughout all of the models was that the specified relations placed domains with easier items as prerequi-

sites for "success" in domains with more difficult items (there were only three specified relations in all of the models considered for which this was not true). This observation tends to suggest that level of item difficulty found within domains plays an important role in determining whether one domain is seen as a prerequisite for another domain.

In order to determine how many conditional relations could "effectively" be specified, comparisons were made among the models with varying numbers of specified conditional relations. This was done separately for both of the empirical generation procedures. Here it was found that, as the number of specified conditional relations were increased, there was simultaneous decrease in the mean  $H_{ij}$  values provided by the models. This suggests that both of these generation procedures tend to specify conditional relations which show less strength as the number of specified relations are increased. This is seen as being a major asset for these procedures, since it would allow an investigator to increase the average strength of specified relations within a model simply by decreasing the number of relations specified.

This phenomenon of decreasing  $H_{ij}$  values which accompanies the increase of specified relations presented a problem for identifying an "optimum" number of specified relations. This is because it was desirable to obtain two antithetical characteristics in a model. First, it was desirable to specify an "adequate" number of conditional relations to clearly describe any existing relations among domains, while at the same time specifying only those conditional relations which were "strong." Thus, depending on how relatively important each of these characteristics is to an investigator, models with differing numbers of conditional relations may be seen as "most desirable."

Further comparisons made among the models with respect to their "added" conditional relations (i.e., those conditional relations found within a given model but not found in similarly generated models with fewer specified relations) tended to suggest that the models with eight out of the thirty possible specified relations provided the most desirable number of conditional relations. This was true for both of the empirical generation procedures. The rationale behind this interpretation was that these particular models allowed for the specification of a maximum number of conditional relations without the occurrence of large decreases in the "added" conditional relations.

It is interesting to note that the "most desirable" models which were generated by the two different empirical generation procedures, are in fact identical models. The most prominent characteristic of these models is their specification of domains 10 and 12 as prerequisites for all of the other domains. This may tend to suggest that the specific skills dealt with in these two domains plays an important role in



preparing students for the acquisition of more difficult content in the general area of multiplication.

Further comparisons were made among models with equivalent numbers of specified relations that were hypothesized on the basis of the different generation procedures. This was done to determine which of the three generation procedures produced the "best fitting" models and thus might be more desirable for use in future research.

The results of these comparisons seemed to suggest that the " $H_{ij}$ " and the "response frequency" generation procedures generated models based on pilot data which provided strikingly similar "fits" to the data in the main study. This finding is quite reasonable if one considers the close similarity of these models. In the case of the models with four and ten specified conditional relations, only two of the specified relations in each case differed from one generation procedure to the other, while in the case of eight specified conditional relations, the models were found to be identical. In general, the comparisons between the two empirical generation procedures tend to suggest that they both provide an effective means of generating hierarchical models. In order of "effectiveness," one might choose the "response frequency" procedure since it provided slightly better "fit" in the case of the models with 10 specified relations. However, in making a choice, it should be noted that the "response frequency" procedure did not allow for the generation of as many different models. It also did not allow for the specification of the exact number of conditional relations desired within a model.

Comparisons made between the "theoretical" hierarchical model for domains based on "experts'" judgments and the corresponding empirical models based on pilot data, suggested that the latter provided a somewhat better fit to the data. This seems to imply that "theoretical" generation may be a less effective means of accurately describing relations among domains. This finding also tends to negate the validity of the structure suggested by the "theoretical" model and raises a question as to how its structure might be improved.

As might be expected, the similarity found between the "theoretical" model and the corresponding "empirical" models were less striking than those found between the two empirical models. A comparison of these models showed that of the ten relations specified by the "theoretical" model, there were, respectively, four and five relations which were inconsistent with those specified in the " $H_{ij}$ " and "response frequency" models. Those portions of the "theoretical" model which were found to be inconsistent with the corresponding empirical models provided a  $\bar{H}_{ij}$  value of .455. However, those portions which were in agreement provided a  $\bar{H}_{ij}$  value of .598. These findings



tend to invalidate those parts of the "theoretical" model which were inconsistent with the empirical models, while lending support to the remainder of the model. It is interesting to note that the conditional relations which are assessed as invalid, deal mainly with the importance of the skills in domain 13 as a prerequisite to the acquisition of other multiplication skills (i.e., domain 13 as a prerequisite for domains 12, 15, and 17 along with domain 10 as a prerequisite for domain 12 were found in the "theoretical" model but not in the corresponding empirical models). Information regarding invalidated parts of the theoretical model are seen as providing valuable information to educators. This is because it allows them to raise questions about why these particular conditional relations did not hold for students and what implications this may have.

One factor which placed limitations on the possible interpretation of the meaning of domain hierarchies was that teaching procedures were not actively manipulated. Thus, it was not possible to determine the extent to which different kinds of variables were actually affecting both the generation and fit of the obtained hierarchies. It is possible that either manner or order of presentation of content could affect the structure of the underlying domain hierarchy. On the other hand, the structure of the underlying hierarchy may be dependent on the manner in which skills related to various domains, form necessary prerequisites for the mastery of items from other domains.

## REFERENCES

- Bart, W. M. and Krus, D. J. An ordering-theoretic method to determine hierarchies among items. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1973, 33, 291-300.
- Gagne, R. M. The acquisition of knowledge. *Psychological Review*, 1962, 69, 355-365.
- Glaser, R. and Nitko, A. Measurement in learning and instruction. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington: American Council on Education, 1971. Pp. 625-670.
- Hively, W. Domain-referenced achievement testing. Paper read at the A.E.R.A. National Convention, 1970.
- Macready, G. B. An investigation into the nature of interitem relations and the structure of domain hierarchies found within a domain referenced testing system. Unpublished doctoral dissertation. University of Minnesota, 1972.
- Macready, G. B. and Merwin, J. C. Homogeneity within item forms in domain referenced testing. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1973, 33, 351-360.

- Patterson, H. L. and Vierling, J. S. *EDINET instruction series: Individualized mathematics program*. Minneapolis: Honeywell Information Systems, 1970.
- Resnick, L. B. and Wang, M. C. Approaches to validation of learning hierarchies. Paper read at the Eighteenth Annual Western Regional Conference on Testing Problems, 1969.

## SOME MULTIPLE RANGE TESTS FOR VARIANCES

KENNETH J. LEVY

State University of New York at Buffalo

Often times, the experimenter is interested in making inferences about treatment variances instead of, or in addition to, inferences about means. Three multiple range tests are proposed for the purpose of specifying which treatment variances or sets of variances are homogeneous. The procedures are based upon the  $F_{\max}$  statistic, Cochran's statistic, and a normalizing log transformation of the sample variances. All three tests depend heavily upon the underlying assumption of normality.

ALTHOUGH interest in means is frequently much greater than interest in variances both theoretically and empirically, nevertheless, instances do arise in which the experimenter is specifically concerned with the problem of making inferences about variances. Various procedures are available for testing the hypothesis of homogeneity of variance associated with  $k$  independent normal populations  $N(\mu_i, \sigma_i^2)$   $i = 1, 2, \dots, k$  with unknown variances  $\sigma_i^2$ ,  $i = 1, 2, \dots, k$ . While the rejection of the hypothesis of equal treatment variances may be statistically interesting, it is in general not very useful. To know simply that a set of treatment variances differ is of limited use, because one still does not know which treatment variances differ from one another. In 1956, H. A. David proposed a multiple range test for variances utilizing the  $F_{\max}$  statistic ( $s_{\max}^2/s_{\min}^2$ ) and Duncan's (1955) philosophy with respect to the choice of significance levels at the various stages of the test. The present paper proposes three different multiple range tests based upon the Newman-Keuls (1939; 1952) philosophy with respect to these significance levels. The three tests will utilize the  $F_{\max}$  statistic, Cochran's statistic ( $s_{\max}^2/\sum_{i=1}^k s_i^2$ ), and a normalizing log transformation of the sample variances respectively.

With respect to a test for variances, the Newman-Keuls and Duncan

procedures may be clearly differentiated by defining a  $p$ -variance significance level. For two variances  $\sigma_1^2, \sigma_2^2$ , let  $D(\sigma_1^2 \neq \sigma_2^2)$  denote the decision  $\sigma_1^2 \neq \sigma_2^2$ . For three variances  $\sigma_1^2, \sigma_2^2, \sigma_3^2$ , let  $D(\sigma_1^2 \neq \sigma_2^2 \cup \sigma_2^2 \neq \sigma_3^2 \cup \sigma_1^2 \neq \sigma_3^2)$  denote the decision that at least one of the variances differs from the other two and they may all be different. For a group of  $k$  variances the two-variance and three-variance significance levels for the sets  $\sigma_1^2, \sigma_2^2$  and  $\sigma_1^2, \sigma_2^2, \sigma_3^2$ , are, respectively,

$$\begin{aligned} \alpha(\sigma_1^2, \sigma_2^2) &= \sup_{\sigma_3^2, \dots, \sigma_k^2} P(D(\sigma_1^2 \neq \sigma_2^2) \mid \sigma_1^2 \\ &= \sigma_2^2; \sigma_3^2, \dots, \sigma_k^2 \text{ arbitrary}) \end{aligned}$$

$$\begin{aligned} \alpha(\sigma_1^2, \sigma_2^2, \sigma_3^2) &= \sup_{\sigma_4^2, \dots, \sigma_k^2} P(D(\sigma_1^2 \neq \sigma_2^2 \cup \sigma_2^2 \neq \sigma_3^2 \cup \sigma_1^2 \neq \\ &\neq \sigma_3^2 \mid \sigma_1^2 = \sigma_2^2 = \sigma_3^2; \sigma_4^2, \dots, \sigma_k^2 \text{ arbitrary})). \end{aligned}$$

Generally, then, a  $p$ -variance significance level for a group of  $p$  variances is the probability of falsely rejecting the hypothesis that the  $p$  variances are all equal, this probability being maximized over the remaining  $k-p$  variances.

For the Newman-Keuls procedure, the  $p$ -variance significance levels for any  $p$  are set equal to  $\alpha$ . In contrast, the Duncan  $p$ -variance significance levels for a given  $p$  are  $1 - (1 - \alpha)^{p-1}$ ,  $p = 2, 3, \dots, k$ . Miller (1966) points out that for means, the Duncan levels do not rise as rapidly as the nonsimultaneous separate test levels; however, they do increase with a fair amount of speed and soon exceed  $1/2$ . Further, he asserts that this violates the spirit of what simultaneous inference is all about, namely, to protect a multiparameter null hypothesis against any false declarations due to the large number of declarations required. For these reasons, this author prefers the Newman-Keuls procedure.

A multiple range test for variances may be performed utilizing the  $F_{\max}$  statistic, Cochran's statistic, or a procedure based upon a log transformation of the sample variances. The  $F_{\max}$  and Cochran's statistics are both discussed at some length in Winer (1971) with respect to general tests for homogeneity of variance. Bartlett and Kendall (1946) investigated a normalizing log transformation of the sample variance  $s^2$  when sampling from a  $N(\mu, \sigma^2)$  population. They showed that  $\log_e s^2$  is approximately normally distributed as  $N(\log_e \sigma^2, 2/n)$  where  $n$  is the number of degrees of freedom for  $s^2$ . This transformation could be used in the following manner as a general test for homogeneity of variance.

Suppose that independent random samples each of size  $n + 1$  are

drawn from  $k$  normal populations  $N(\mu_i, \sigma_i^2)$ ,  $i = 1, 2, \dots, k$  with unknown means and variances. When the null hypothesis is true, i.e., when  $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 = \sigma^2$ , then the  $\log_e s_i^2$  will be independent and identically distributed, approximately as  $N(\log_e \sigma^2, 2/n)$ . Pearson and Hartley (1942), have obtained the probability integral for the following statistic:

$$R_n = \frac{X_{\max} - X_{\min}}{\sigma}$$

where  $X_{\max} - X_{\min}$  is the range of an independent random sample of size  $n$  drawn from a  $N(\mu, \sigma^2)$  population. Thus, when the null hypothesis is true for  $k$  groups, then

$$\frac{\log_e s_{\max}^2 - \log_e s_{\min}^2}{\sqrt{2/n}}$$

will be approximately distributed as  $R_k$ . If the observed value of  $R_k$  exceeds the critical value obtained from the Person and Hartley table, one would reject the hypothesis of homogeneity of variance.

Three multiple range tests will now be performed upon a hypothetical set of data following a procedure for testing means outlined in Winer (1971).

Suppose that a completely randomized experiment with 4 treatments and 10 subjects per treatment has been conducted. The experimenter wishes to make inferences concerning which treatment conditions are homogeneous with respect to variance.

1. Compute the sample variances of the 10 measures in each of the 4 experimental groups.

$$s_i^2 = \sum_{j=1}^n \frac{(x_{ij} - \bar{x}_i)^2}{n-1}$$

where  $n$  = the number of subjects in the  $i$ th group and  $n-1$  degrees of freedom for  $s_i^2$ .

Suppose the following results were obtained:

Group	$s^2$
1	.75
2	12.00
3	3.00
4	2.00

2. Order the sample variances horizontally and vertically as follows:

	(1)	(4)	(3)	(2)
	.75	2.00	3.00	12.00
(1)	.75			
(4)	2.00			
(3)	3.00			

(1) denotes that .75 is the sample variance associated with group 1.

3. Compute the  $F_{\max}$  statistic for 4 groups taking the ratio of the extremes (2)/(1) i.e.,  $F_{\max} = 12.00/.75 = 16.00$ . From tables in Winer



(1971), the critical value for a .05-level test for 4 groups each with 9 degrees of freedom for  $s^2$  is 6.31. Since the observed ratio exceeds 6.31, the hypothesis of homogeneity of variance is rejected. Therefore, an asterisk should be entered in cell (1, 2) in the above table.

The basic credo of this multiple range test may be stated as follows: the differences between any two variances in a set of  $k$  variances is significant provided the ratio of the extremes (max/min) of each and every subset which contains the given variances is significant according to an  $\alpha_p$  level test where  $p$  is the number of variances in the subset concerned. Thus, if the initial test ratio were not significant, no further tests would be made.

4. Since the initial observed ratio was significant, proceed to test the diagonal ratios (3)/(1) and (2)/(4). The critical value for a .05-level test for 3 groups each with 9 degrees of freedom for  $s^2$  is 5.34.  $F_{\max} = (3)/(1) = 3.00/.75 = 4.00$ ;  $F_{\max} = (2)/(4) = 12.00/2.00 = 6.00$ . Thus, do not reject the hypothesis that  $\sigma_1^2 = \sigma_4^2 = \sigma_3^2$ ; however, do reject the hypothesis that  $\sigma_4^2 = \sigma_3^2 = s^2$ . Since the ratio (3)/(1) is not significant, one is not allowed to test any other differences within the triangle bounded in the upper right corner by the cell (1, 3). Since the ratio (2)/(4) is significant, an asterisk is recorded in cell (4, 2) and one proceeds to test the ratio (2)/(3).

5.  $F_{\max} = (2)/(3) = 12.00/3.00 = 4.00$ . The critical value for a .05-level test for 2 groups each with 9 degrees of freedom for  $s^2$  is 4.03. Since the observed ratio does not exceed the critical value, do not reject the hypothesis that  $\sigma_3^2 = \sigma_2^2$ .

In summary, then, one may conclude that the variances associated with groups 1, 4, and 3 do not differ; groups 3 and 2 do not differ; however, the variance of group 2 is significantly greater than the variance of either group 1 or group 4. In terms of Duncan's notation:

$$\begin{array}{cccc} (1) & (4) & (3) & (2) \\ \hline & & & \end{array}$$

where groups underlined by a common line do not differ from one another; groups not underlined by a common line do differ.

Similar tests may be performed utilizing Cochran's statistic and the normalizing log transformation discussed above. For a test based upon Cochran's statistic, the initial observed ratio would be

$$\begin{aligned} C &= \frac{s_{\max}^2}{\sum s_i^2} i = 1, 2, \dots, k \\ &= \frac{12.00}{(.75 + 2.00 + 3.00 + 12.00)} \\ &= .6761 \end{aligned}$$

From tables in Winer (1971), the critical value for a .05-level test for 4 groups each with 9 degrees of freedom for  $s^2$  is .5017. Since the observed ratio exceeds .5017, reject the hypothesis of homogeneity of variance and proceed to step 4.

For a test based upon the normalizing log transformation, one must compute  $\log_e s_i^2$ .

Group	$s^2$	$\log_e s^2$
1	.75	-.2877
2	12.00	2.4849
3	3.00	1.0986
4	2.00	0.6931

The initial test statistic for this case is:

$$R_4 = \frac{(\log_e s_{\max}^2 - \log_e s_{\min}^2) \sqrt{n-1}}{\sqrt{2}} = \frac{(2.4849 + .2877)3}{\sqrt{2}} = 5.8816$$

From the Pearson and Hartley (1942) tables, the critical value for a .05-level test based upon the range of the  $\log_e s_i^2$  for 4 groups is 3.62. Since the observed ratio exceeds 3.62, again reject the hypothesis of homogeneity of variance and proceed to step 4.

It should be noted, that in general, tests for homogeneity of variance are extremely sensitive to violations of the underlying assumption of normality. Box (1953) points out that most tests do not utilize any evidence of variance variability within the samples. The sample variability is measured theoretically, and the theoretical variability changes as the underlying distribution changes. For this reason, the above tests were studied via Monte Carlo techniques with sampling occurring from normal, uniform, and double exponential populations. It was found that all three tests were seriously affected by non-normality with respect to both their significance levels and their empirical power. When sampling occurred from normal populations, the  $F_{\max}$  test and the test based upon the normalizing log transformation were found to be most sensitive to those cases in which a sample variance was anomalously small. In contrast, the test based upon Cochran's statistic was found to be most sensitive to those cases in which one or more sample variances were anomalously large. In conclusion then, one should be reasonably confident about the assumption of normality before proceeding with any of the procedures discussed above.

## REFERENCES

- Bartlett, M. S. and Kendall, D. G. The statistical analysis of variance heterogeneity and the logarithmic transformation. *Royal Statistical Society*, 1946, 8, 128-138.
- Box, G. E. P. Non-normality and tests on variances. *Biometrika*, 1953, 40, 318-335.
- David, H. A. The ranking of variances in normal populations. *Journal of the American Statistical Association*, 1956, 51, 621-626.
- Keuls, M. The use of the "studentized range" in connection with an analysis of variance. *Euphytica*, 1952, 1, 112-122.
- Miller, R. G. *Simultaneous statistical inference*. New York: McGraw-Hill, 1966.
- Newman, D. The distribution of the range in samples from a normal population, expressed in terms of an independent estimate of standard deviation. *Biometrika*, 1939, 31, 20-30.
- Pearson, E. S. and Hartley, H. O. The probability integral of the range in samples of  $n$  observations from a normal population. *Biometrika*, 1942, 32, 301-310.
- Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1971.

## JUDGMENTAL BIAS IN THE RATING OF ATTITUDE STATEMENTS<sup>1</sup>

WILLIAM H. BRUVOLD<sup>2</sup>

University of California, Berkeley

Judges holding divergent attitudes toward high contact uses of water reclaimed from community sewage rated two sets of attitude statements regarding this issue. Results showed a close linear relationship between item scale values obtained from positive and negative attitudinal groups, and also a somewhat reduced range of ratings for judges holding unfavorable personal attitudes toward reuse. These findings, and the findings of previous research on this issue, were seen as being consonant with an item displacement theory of rating performance and supportive of equal interval measurement.

CONTROVERSY has encompassed the equal-appearing intervals attitude scaling procedure ever since Hovland and Sherif (1952) challenged the assumption that judges' personal attitudes do not influence item placement. Of the many attempts to deal with the issue, five relatively distinct explanations of the influence of attitude upon the judgmental process seem to be most prominent at this time.

First, the well known notions of Hovland and Sherif (1952) continue to maintain their importance primarily because these authors were the first to clearly stipulate that relationships between median item placements would be specifiably curvilinear for different attitudinal groups. The curvilinear hypothesis has important implications for a theory of judgment and important implications for measurement theory. If median item placements are not linearly related then attitude scales con-

---

<sup>1</sup> This research was supported by grant W-359 from the University of California Water Resources Center.

<sup>2</sup> Requests for reprints should be sent to William H. Bruvold, School of Public Health, University of California, Berkeley, California 94720.

structed by the method of equal-appearing intervals can not represent interval level measurement (Torgerson, 1958).

Second, and in response to the Hovland and Sherif (1952) paper, Hinckley (1963) replicated his earlier work supposedly demonstrating no influence of attitude on item rating. The no influence position was reasserted on the basis of the replication study. The implication of this position regarding interval level measurement is very clear. If judges' personal attitudes do not influence item ratings, then median item placements for any two groups of judges, regardless of their typical attitude, should be identical. Such a result would produce precisely linear relationships between item medians having slope one and intercept zero. Such results would support, rather than refute, attainment of interval level measurement (Torgerson, 1958).

Third, Upshaw (1965) has developed a variable series theory maintaining the linearity notion contained in the no influence of attitude position and also accommodating the idea that attitudes do have a systematic effect on ratings. The outcome of Upshaw's (1965) effort is that interval level measurement is supported in terms of Torgerson's (1958) master scale evaluation technique, yet attitude is seen as an important determinant of obtained judgments through its influence over the judge's perspective on the items rated. The linearity aspect of Upshaw's (1965) expectation was reasonably validated by his own research while the expectation regarding the influence of attitude on median item ratings was not confirmed.

In fact Upshaw's (1965) findings on the relation of attitude to median item placement were consistent with the findings of Zavalloni and Cook (1965). The latter note that their results appear to confirm an emerging generalization about the relation of judges' attitudes and item placement on the equal-appearing interval continuum; namely, that judges with favorable attitudes tend to employ a wider range of ratings which results from more positive ratings for positive items, and more negative ratings for negative items, than those given by judges having unfavorable attitudes toward the matter at hand. Zavalloni and Cook (1965) account for these results in terms of the judge's agreement or disagreement with items in conjunction with a tendency to rate agreed-with items more favorably and disagreed-with items less favorably. While these authors posit a clear and definite effect of attitude upon judgments, the implication of their explanation for measurement theory is not clear. The displacement theory of Zavalloni and Cook (1965) theory could remain viable with either linear or curvilinear relations between median item placements; however, it appears that displacement theory would expect near linear relationships if certain neutral items show little displacement. If relationships are



linear, or nearly so, and if attitudes of the more favorable group were represented on the  $x$ -axis, the slope should be less than one and the  $y$ -intercept greater than zero.

The fifth prominent explanation of the influence of attitudes upon item judgment has recently been developed by Eiser (1971). This effort appears to be primarily an elaboration of the Zavalloni and Cook (1965) position. The major effect of attitude is held to be item displacement as described above and it is supposedly due to perceived item contrast based upon the judges' personal reaction to individual items. This effect does not operate alone, since, in addition, measureable effects of social norms and anchoring are posited. The joint operation of these three factors is supposed to produce median items for different attitudinal groups that are curvilinearly related. Eiser (1971) does not specify the character of the nonlinearity expected nor its basis in the three factors posited. Nevertheless, the expectation of curvilinearity represents an extension of the Zavalloni and Cook (1965) explanation that is uncongenial to interval measurement.

The character of the mathematical relationship between median item placements obtained from divergent attitudinal groups is a central issue for the five explanations here reviewed. Two of the five predict curvilinearity, two predict linearity, and one suggests a certain kind of linear relationship. Further, as mentioned above, the presence or absence of linearity is most important for reasons involving theory of measurement. In view of the centrality of this issue it is surprising to note that none of the five articles referenced used standard statistical procedures (McNemar, 1969) to describe and evaluate the relationship of interest. Analyses have involved either course groupings and analysis of variance without tests for trends, or simple correlation coefficients without accompanying  $\eta$  values. A major purpose of the present study is to assess the relationship of interest using appropriate statistical techniques that will more adequately evaluate the five competing explanations here summarized and their implications for theory of measurement.

### *Method*

One hundred and seven individuals served voluntarily as judges in this research and all were residents of California. Sixty-four judges were male and ages ranged from 26 to 73 years. An attempt was made to obtain the widest possible spectrum of opinion on the matter of reuse of reclaimed water and therefore judges were recruited from University classes, a professional engineering group, a well known conservation society and a local social club. No strict sampling plan

was followed, rather the general aim of recruitment was to obtain a diverse group of judges.

Two groups of statements, developed in an earlier and totally separate study (Bruvold, 1971), dealing with the use of water reclaimed from sewage for drinking, or for swimming, were employed in this research. There were 83 statements regarding drinking and 58 involving swimming. It should be strongly emphasized that statements regarding beliefs or behavioral intentions were deliberately excluded from the item pool.

Items were typed, one to a plain white 3" × 5" card, and labeled at random with a one letter-three digit number combination which appeared in the upper right corner of each card. Order of items within sets, and order of the two sets themselves, were randomly arranged before each individual judging session. Standard equal-appearing interval judgment procedures were used as in the earlier work (Bruvold, 1971). Each judge rated, for practice only, under the observation of the experimenter, five statements not included in the major item sets. Questions regarding judgment procedures were fully discussed before proceeding to the major item sets. Each judge worked alone to complete the major task, each gave a complete set of ratings, and the ratings of all were included in subsequent statistical analyses.

Upon completion of the item rating task each judge's personal attitude toward reclaimed water for drinking, and swimming, was assessed by two Thurstone scales. Each scale was comprised of twenty statements selected from Remmers (1934) stems dealing with attitude toward any practice. Reclaimed water for drinking, or for swimming, was inserted in place of the term "this practice" in each statement. No item was common to both scales. In completing each scale the judge was required to check the three or four items nearest to his own personal attitude. Scores on both scales could range from a low of 1.0 to a high of 11.0.

### *Analysis and Results*

Attitude scale scores were obtained for each judge and scale and then rank ordered separately by scale. Scores for the drinking scale ranged from 2.0 to 9.9 and those for swimming from 2.2 to 9.7. As in other work the distributions were divided into quintiles here containing 21 judges each except for the third which contained 23 judges. Remaining analysis focus upon the lowest and highest quintiles. For drinking the highest attitude score in the first quintile was 4.2 and the lowest score in the fifth quintile was 8.6. Analogous figures for swimming were 4.8 and 8.4. Four sets of equal-appearing interval

item scale values were obtained by standard methods (Bruvold, 1971), one for each scale and attitudinal position. Subsequently, the relationship between median item scale values obtained from judges positive or negative in attitude toward reclaimed water for drinking was analyzed using statistical tests for linearity and curvilinearity proposed by McNemar (1969). The same procedure was also performed on the swimming data. These analyses were made possible by taking all median item values from the positive judges, the independent or  $x$ -variable, as category mid-points of 1.5, 2.5, . . . , 10.5 while median item values from the negative judges, the dependent or  $y$ -variable, were left ungrouped. Thus, for example, any item scale value from the positive attitude group beginning with the integer 6 was recorded as 6.5 whereas associated  $y$ -variable item values from the negative attitude group were left unchanged.

Results from the swimming data are now summarized. These analyses were performed for 83 items and the regression equation was  $y' = 0.944x + 0.866$ . The  $F$ -ratio for the linear component of the regression was 2,199.98 ( $p < .001$ , 1/73  $df$ ) and it was 1.73 ( $p > .05$ , 8/73  $df$ ) for the curvilinear component. The Pearson  $r$  between positive and negative item values was 0.981 and the corresponding eta equalled 0.984. A  $t$  test assessing the deviation of the obtained slope from 1.000 equalled 2.80 ( $p < .01$ , 81  $df$ ). Analogous figures for the 58 swimming items are now presented. The regression equation was  $y'' = 0.889x' + 0.985$ . The  $F$ -ratio for linearity was 1,764.61 ( $p < .001$ , 1/48  $df$ ) and for curvilinearity it was 0.178 ( $p > .05$ , 8/48  $df$ ). Pearson  $r$  equalled 0.985 and eta was 0.988. The  $t$  test of the difference of the obtained slope from unity was 5.55 ( $p < .001$ , 56  $df$ ).

### *Discussion*

The results here obtained were most uncongenial to positions (Hovland and Sherif, 1952; Eiser, 1971) predicting curvilinear relationships between median item ratings obtained from groups divergent in personal attitude toward the issue under study. Results of the regression analyses reported above showed that curvilinearity failed to reach customary criteria for statistical significance. This interpretation is strongly substantiated by comparison of  $r$  and eta values. Thus expectations predicting systematic non-linearity receive little support from these data and, further, the results do not indicate that interval level measurement is precluded for equal-appearing interval methods because of the effects of personal attitudes upon judgment-produced scales.

Absence of significant curvilinearity between median item ratings

for divergent attitudinal groups, does not give Hinckley's (1963) no-effect position the support of these data. As noted above, Upshaw (1965) obtained an item displacement effect, more fully documented by Zavalloni and Cook (1965) and then confirmed by Eiser (1971) which indicates that judges holding unfavorable personal attitudes rate negative items more positively, and positive items more negatively, than judges holding favorable personal attitudes. The general implications of this expectation for linear slope and intercept values were outlined in the introduction section. It may be noted here that both regression equations reported above fit these general expectations. The analysis of the difference of obtained slopes from unity shows that these observed values were significantly less than one. Somewhat larger rating dispersions obtained for the positive attitudinal groups are consonant with the slope results. The consistency of these findings yields additional support for the Zavalloni and Cook (1965) position.

Summarizing, it may be stated that the present data provide support for linear item displacement theory emerging in the work of Upshaw (1965), and Zavalloni and Cook (1965). Item displacement theory need not invoke apparently complex explanatory constructs. Rather, the major effects of item displacement may be due to competing tendencies faced by an individual when learning to perform the judgments required by the method of equal-appearing intervals. The writer and his students have noticed that many individuals experience difficulty when learning to give equal-appearing interval ratings. Most, if not all, do not seem initially to understand that their judgments of, and not responses to, items are sought. The difficulty may stem from the usual psychometric practice of asking for responses to items rather than judgments. If raters do not, or can not, fully set aside a response tendency, then the rating of a particular item will be the result of two factors: how well the rater likes and agrees with the item, and also his judgment of where it belongs on the equal-appearing interval continuum. Consonant with the notions of Zavalloni and Cook (1965), the joint operation of competing response and judgmental factors would result in smaller rating dispersions, higher ratings for negative items, and lower ratings for positive items, for raters unfavorable in personal attitude when compared to raters holding favorable personal attitudes toward the issue at hand.

The position here expressed, while seeming to best account for past and present results, also suggests topics for future study. First, item displacement would likely be reduced as adequacy of rating instruction increases. Since the present study used reasonably thorough instructions involving several example statements, the item displacement



effect, it would follow, should not have been large. Future research can determine if this reasoning has predictive value. Second, item displacement would likely be reduced as rating instructions are understood and accepted. Third, demand characteristics or perceived social pressure to respond to items rather than to judge them would likely enhance item displacement. Reading of the Hovland and Sherif (1952) article suggests that such effects may have been operative in that research. Conversely, perceived social pressure not to respond to items, but to try to judge them "objectively," would likely reduce item displacement. Finally, analysis of the tendency to personally respond to individual items should prove interesting. Such analysis could lead to further expectations regarding individuals and items most likely to evidence displacement.

## REFERENCES

- Bruvold, W. H. Affective response toward uses of reclaimed water. *Journal of Applied Psychology*, 1971, 55, 28-33.
- Eiser, J. R. Enhancement of contrast in the absolute judgment of attitude statements. *Journal of Personality and Social Psychology*, 1971, 17, 1-10.
- Hinckley, E. D. A follow-up study on the influence of individual opinion on the construction of an attitude scale. *Journal of Abnormal and Social Psychology*, 1963, 67, 290-292.
- Hovland, C. I. and Sherif, M. Judgmental phenomena and scales of attitude measurement: Item displacement in Thurstone scales. *Journal of Abnormal and Social Psychology*, 1952, 47, 822-832.
- McNemar, Q. *Psychological Statistics* (4th ed.). New York: Wiley, 1969.
- Remmers, H. H. Studies in attitudes: A contribution to social-psychological research methods. *Bulletin of Purdue University*, 1934, 35, 64-67.
- Torgerson, W. S. *Theory and methods of scaling*. New York: Wiley, 1958.
- Upshaw, H. S. The effect of variable perspectives on judgments of opinion statements for Thurstone scales: Equal-appearing intervals. *Journal of Personality and Social Psychology*, 1965, 2, 60-69.
- Zavalloni, M. and Cook, S. W. Influence of judges' attitudes on ratings of favorableness of statements about a social group. *Journal of Personality and Social Psychology*, 1965, 1, 43-54.





## EMPIRICAL OPTION WEIGHTING WITH A CORRECTION FOR GUESSING<sup>1</sup>

RICHARD R. REILLY  
Educational Testing Service

Because previous reports have suggested that the lowered validity of tests scored with empirical option weights might be explained by a capitalization of the keying procedures on omitting tendencies, a procedure was devised to key options empirically with a "correction-for-guessing" constraint. Use of the new procedure with Graduate Record Examinations (GRE) data resulted in smaller increases in reliability than those observed when unconstrained procedures were used, but validities for quantitative subforms were not appreciably lowered. Validities for verbal subforms were lowered slightly, however.

Two recent reports (Hendrickson, 1971; Reilly and Jackson, 1972) have suggested that weighting options empirically results in substantial increases in reliability and test homogeneity, but at the expense of lowered test validity. These findings are at variance with those reported in an earlier study by Davis and Fifer (1959) who found similar increases in reliability and slight increases in validity when options were weighted empirically. All three studies employed modifications of a weighting technique originally known as The Method of Reciprocal Averages (Mosier, 1946) which, in effect, maximizes the product-moment correlation between item scores and criterion scores by assigning to each item-option values proportional to the mean criterion score for all individuals choosing that option.

A key difference between the Davis and Fifer study and the first two mentioned was that tests in the first two were administered with formula score instructions while Davis and Fifer instructed examinees to

---

<sup>1</sup> The research reported herein was supported by the Graduate Record Examinations Board.

attempt every item. Thus, Hendrickson and Reilly and Jackson had an additional "option," that of omit. Hendrickson, reporting on the weights generally assigned to the omit category comments, "... An interesting finding of this study was that the weight of 'omit' was almost always lower than any of the other distracters in an item..." (Hendrickson, 1971). Reilly and Jackson (1972) take this a step further and suggest that, "... the empirical keying procedures described capitalize on the tendency to omit and ... while this tendency is reliable, it is not valid."

Because of these suggestions, it was decided to devise and test a procedure which weighted options subject to the constraint that the weight for omit equal the mean weight for the options. The rationale is similar to that used in the usual formula scoring method in that it assumes that an individual omitting an item should receive the expected weight under conditions of random response to that item.

In order to determine the optimum weights for a single item, subject to the "correction-for-guessing" constraint, the following objective function was set up:

$$F = \sum_i \sum_j (y_{ij} - w_j)^2 - 2\lambda[(k-1)w_p - \sum_j \delta_j w_j],$$

where

$y_{ij}$  denotes the criterion score of the  $i$ th individual making the  $j$ th response;

$w_j$  is the weight for the  $j$ th response,

$j = 1 \dots p, \dots k$ ; and

$w_p$  is the weight for the omit category.

$\delta_j =$  one for  $j \neq p$ , and zero otherwise; and

$\lambda$  is the LaGrange multiplier.

Taking partial derivatives and solving for the weights which minimize the function we find that the solution, which requires a small  $[(k-1) \times (k-1)]$  matrix inversion, has the following properties<sup>2</sup>: (1) The mean item score over all individuals is equal to the mean criterion score; (2) the weights arrived at are proportional to the weights which will maximize the correlation between the item and the criterion subject to the constraint of a fixed item variance (and, of course, the constraint that the omit weight equals the mean of the option weights); (3) unlike the constrained option weights, the weights arrived at will not, in general, yield the maximum possible product-moment correlation; (4) for unconstrained weights it has been pointed out (Stanley and Wang, 1970) that a slope of 1.0 and a zero intercept will describe

<sup>2</sup> The full proof is available from the author on request.

the regression of the criterion scores on the item scores. The appropriate slope for the regression of criterion scores on item scores yielded by the new method will not, in general, be 1.0, nor will the appropriate intercept, in general, be zero.

### Procedure

Two parallel forms each, of the verbal (denoted as  $V_1$  and  $V_2$ ) and quantitative ( $Q_1$  and  $Q_2$ ) sections of the Graduate Record Examinations (GRE), were devised by assigning one-half of the items on each section to each of the two special parallel forms. Forms  $V_1$  and  $V_2$  consisted of 50 items each, while forms  $Q_1$  and  $Q_2$  consisted of 27 items each. It should be noted that the two forms in each set, since they were constructed from operation tests, were not administered under separate time limits. Because of practical limitations the more desirable procedure of administering the two parallel forms under separately timed conditions was not possible.

Data were the same as these used in the Reilly and Jackson (1972) study. A spaced sample (i.e., a sample consisting of every  $n$ th answer sheet) of 5,000 answer sheets (sample A) from the December 1970 administration of the GRE was employed for study purposes. A second sample (sample B) consisting of the answer sheets of 4,916 individuals from the same administration was taken for validation purposes. Sample A was divided into two randomized block groups of 2,500 (samples  $A_1$  and  $A_2$ ) by blocking on total GRE score. The 5,000 answer sheets were ordered in terms of the verbal score plus the quantitative score and then randomly assigned to the two subsamples. This increased the likelihood that the two split samples would be comparable in terms of total score distributions. Each subtest was keyed against the scores on its parallel form in sample  $A_1$ . The tests in sample  $A_2$  were then scored using these derived weights and intercorrelations, and alpha coeffi-

TABLE 1  
*Cross-Validated Parallel Forms Reliabilities for  
Empirically Keyed and Formula Scored Subtests*

	Formula	Empirically Keyed	$K^a$
Verbal	.8909	.9242	1.49
Quantitative	.8742	.8892	1.16

<sup>a</sup>  $K$  gives the estimated proportional increase in test length which would be necessary to yield the increased  $R$ s shown. Rearranging the Spearman-Brown prophecy formula,

$$K = \frac{R_w(1 - R_f)}{R_f(1 - R_w)}$$

where  $R_f$  is the  $R$  obtained with formula score weights and  $R_w$  is the cross-validated  $R$  obtained with empirical weights.

TABLE 2  
*Cross-Validated Internal Consistency Coefficients for  
 Formula Scored and Empirically Keyed Tests*

	Formula	Empirically Keyed	K <sup>a</sup>
V <sub>1</sub>	.8745	.9069	1.40
V <sub>2</sub>	.8755	.9084	1.41
Q <sub>1</sub>	.8515	.8817	1.30
Q <sub>2</sub>	.8725	.8852	1.13

<sup>a</sup> K gives the estimated proportional increase in test length which would be necessary to yield the increased  $\alpha$ 's shown. Rearranging the Spearman-Brown prophecy formula,

$$K = \frac{\alpha_w(1 - \alpha_F)}{\alpha_F(1 - \alpha_w)}$$

where  $\alpha_F$  is the  $\alpha$  obtained with formula score weights and  $\alpha_w$  is the cross-validated  $\alpha$  obtained with empirical weights.

cients were computed. Thus, all results reported are those obtained with cross-validated weights.

The next step involved scoring the sample B answer sheets and computing the single order and multiple correlations between the empirically keyed tests and *undergraduate* GPA. Sample B was drawn from a total of 40 different colleges. Within-school samples ranged from a low of 16 to a high of 399. A modification of one of Tucker's (1963) central prediction methods was used to pool data across colleges.<sup>3</sup>

### *Results and Discussion*

The results of the keying on parallel forms reliability and internal consistency are presented in Tables 1 and 2. The proportional increases in effective test lengths are comparable to those reported by Hendrickson (1971) but less than those observed by Reilly and Jackson (1972). The smaller increments observed for the quantitative tests are consistent with previous findings, and may, as Hendrickson (1971) suggests, be related to the common observation that differences in the quality of the distracters are less apparent for general mathematical items than for verbal items.

Reilly and Jackson (1972) observed increases in the correlations between verbal and quantitative tests when empirical weights were used and attributed these increases to the capitalization of the keying procedure on an omitting factor common to both tests. Thus, the results shown in Table 3 are of interest since they indicate that when constrained weights are used the large increases in verbal-quantitative

<sup>3</sup> The method used is a least-squares procedure worked out by Robert F. Boldt and is more fully described in a report by Briggs (1970).



TABLE 3  
*Intercorrelations between Verbal and Quantitative Forms  
 for Formula Scored and Empirically Keyed Tests*

	Formula	Empirically Keyed	Expected <sup>a</sup>
V <sub>1</sub> Q <sub>1</sub>	.4154	.4577	.4269
V <sub>2</sub> Q <sub>1</sub>	.4190	.4428	.4550
V <sub>1</sub> Q <sub>2</sub>	.4079	.4304	.4191
V <sub>2</sub> Q <sub>2</sub>	.4061	.4138	.4173

<sup>a</sup> The expected values represent the expected correlation which should have resulted from the increased reliability of the empirical key scores. These values were obtained by multiplying the true formula score correlations between V and Q by the geometric mean of the empirical key score reliabilities. Parallel forms reliabilities were used in all cases.

correlations do not occur. When increases in reliability are taken into account the increases are actually slightly less than expected in two of the four cases shown and slightly greater than expected in the remaining two cases.

In Table 4, the correlations are shown between pairs of parallel subtests, one scored with empirical weights and the other with formula weights. These latter correlations are, in general, slightly higher than the parallel forms reliability, in contrast to the uniformly lower values obtained when unconstrained weights were used (Reilly and Jackson, 1972).

The validity results are presented in Table 5. From a conceptual point of view the most desirable criterion for assessing validity would have been some measure of graduate school performance. The small within-school sample sizes as well as the generally restricted variance in graduate grades made this unfeasible. Undergraduate grades are a readily available concurrent measure of academic achievement and seemed a reasonable criterion against which to validate the different scoring methods. While the zero-order validities for the quantitative forms are almost unchanged, the multiple correlations are slightly

TABLE 4  
*Intercorrelations between Empirically Keyed and Formula Scored  
 Parallel Forms*

	Parallel Forms Reliability	Empirically Keyed vs. Formula Scored Parallel Form <sup>a</sup>	
		I	II
Verbal	.8909	.8953	.8914
Quantitative	.8742	.8726	.8848

<sup>a</sup> Column I shows the correlation between form V<sub>1</sub> (Q<sub>1</sub>) empirically keyed and form V<sub>1</sub> (Q<sub>2</sub>) formula scored. Column 2 shows the correlation between V<sub>2</sub> (Q<sub>2</sub>) empirically keyed and V<sub>1</sub> (Q<sub>1</sub>) formula scored.

TABLE 5  
Pooled and Median Validity Coefficients

	Formula Scores	Unconstrained Weighted Scores <sup>a</sup>	Constrained Weighted Scores
V <sub>1</sub> Median	.3069	.2467	.2768
V <sub>1</sub> Pooled <sup>b</sup>	.2703	.3167	.2998
Q <sub>1</sub> Median	.1407	.1299	.1386
Q <sub>1</sub> Pooled	.1664	.1909	.1894
V <sub>1</sub> Q <sub>1</sub> Median	.2768	.3443	.3145
V <sub>1</sub> Q <sub>1</sub> Pooled	.2666	.3184	.2997
V <sub>2</sub> Median	.2987	.2358	.2841
V <sub>2</sub> Pooled	.2939	.2532	.2828
Q <sub>2</sub> Median	.1679	.1504	.1681
Q <sub>2</sub> Pooled	.2055	.1847	.2054
V <sub>2</sub> Q <sub>2</sub> Median	.3135	.2589	.3036
V <sub>2</sub> Q <sub>2</sub> Pooled	.3013	.2637	.2919

<sup>a</sup> The unconstrained weights were those obtained by keying against parallel forms (Reilly and Jackson, 1972).

<sup>b</sup> Pooled single order coefficients were estimated as follows:

$$r = \sqrt{\frac{\sum n_i r_i^2}{\sum n_i}}$$

multiple correlation coefficients were obtained using a pooling procedure described by Briggs (1970).

lower overall owing primarily to the decreases in the correlations between GPA and the empirically keyed verbal subtests. It is difficult to explain why, even with the modified keying procedure, the verbal test validities were lowered. Apparently, the empirically keyed verbal tests are measuring some additional factors which, though reliable, may not be valid.

### Conclusions

While the results reported here certainly do not indicate that steps should be taken to implement empirical option weighting, the findings are not entirely discouraging either. It has been shown that a test can be made more reliable and more homogeneous through option weighting and, at least for the quantitative forms, without any appreciable lowering of validity.

Further research should be done on several key issues which have emerged in this study. First, the issue of omitting behavior should be looked at more closely. Breen (1972) has presented data for the SAT which indicate that "omit" scores are even more reliable than rights-only or formula scores. It may be that an omitting score can be used as a suppressor variable along with the formula score to increase the correlation with the criterion.

Another interesting and potentially useful study would be one which examined the effects of keying options directly on the GPA criterion.

Examination of the weights for options may reveal consistent patterns which could be helpful in guiding item writers.

## REFERENCES

- Briggs, B. Boldt's special case of central prediction, weighted least squares procedure. Statistical Systems Report, SS12. Princeton, N.J.: Educational Testing Service, 1970.
- Davis, F. B. and Fifer, B. The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1959, 19, 159-170.
- Green, B. F. The sensitivity of Guttman weights. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois, April 7, 1972.
- Hendrickson, G. F. The effect of differential option weighting on multiple-choice objective tests. Report No. 93. Center for the study of social organization of schools, The Johns Hopkins University, Baltimore, Maryland, 1971.
- Mosier, C. I. Machine methods in scalling by reciprocal averages. *Proceedings, Research Forum*. New York: International Business Machines Corporation, 1946. Pp. 35-39.
- Reilly, R. R. and Jackson, R. Effects of empirical option weighting on validity and reliability of an academic aptitude test. *Journal of Educational Measurement*, 1972.
- Stanley, J. C. and Wang, M. D. Weighting test items and test-item options, an overview of the analytical and empirical literature. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1970, 30, 21-25.
- Tucker, L. R. Formal models for a central prediction system. *Psychometric Monograph No. 10*. Richmond, Va.: William Byrd Press, 1963.



## THE CONCEPT OF EFFICIENCY IN ITEM ANALYSIS

RICHARD J. HOFMANN

Miami University<sup>1</sup>

In this paper a new item analysis index,  $e$ , is derived as a function of difficulty and discrimination to represent item efficiency. It is demonstrated algebraically that the maximum discriminating power of an item may be determined from its difficulty and then item efficiency is defined as the ratio of observed discrimination to maximum discrimination. The magnitude of the  $e$ -index will range from zero to unity and will provide additional information for item analyses.

In a typical analysis of a test item, two indices are usually computed, a difficulty index and a discrimination index. If one assumes an analysis based upon the performance of two groups on the item, typically referred to as a U-L analysis, then a two by two contingency table may be used in the tabulation of the indices. Such an approach would typify the approaches suggested by Kelley (1939), Johnson (1951) and Cureton (1959) and is discussed in almost any basic measurement text devoting some space to item analyses.

Assume that  $N'$  individuals have responded to some item in either a positive fashion,  $r$ , or a negative fashion,  $w$ . Furthermore, assume that either on the basis of their total scores on the instrument associated with the item or on the basis of some outside criterion two equal groups,  $g_1$  and  $g_2$ , are determined from  $N'$ . In this case  $g_1$  and  $g_2$  in total represent  $N$  individuals where  $N$  may be equal to or proportionate to  $N'$ . Then using the subscripts 1 and 2 to denote those symbols associated with  $g_1$  and  $g_2$ , respectively, the  $N$  responses to the item are presented in Table 1.

---

<sup>1</sup> This research was partially supported by a small grant from the Faculty Research Committee of Miami University.



TABLE 1  
*Contingency Table Summarization of Two Groups Responses to a Single Item*

Response Category	Group		Total
	$g_1$	$g_2$	
Positive	$r_1$	$r_2$	$r_1 + r_2$
Negative	$w_1$	$w_2$	$w_1 + w_2$
Total	$r_1 + w_1$	$r_2 + w_2$	$N$

The difficulty index  $a$  of the item may be denoted as:

$$a = \frac{r_1 + r_2}{N} \quad (1)$$

and it may range from zero to unity. One interpretation of  $a$  is that it represents the proportion of  $N$  individuals responding positively to the item. A second interpretation is that  $a$  represents the probability of observing a positive response to the item,  $P(r)$ .

The discrimination index of the item,  $b$ , may be denoted as:

$$b = \frac{r_1}{n} - \frac{r_2}{n} \quad (2)$$

where

$$r_1 + w_1 + r_2 + w_2 = N$$

and it may range from positive to negative unity. One interpretation of  $b$  is that it represents the difference between two conditional probabilities, the probability of a correct response given membership in group one, less the probability of a correct response given membership in group two.

In a very special sense the marginals of Table 1 are fixed and there is an interdependence between difficulty and discrimination. In this paper discrimination is assumed to be a function of difficulty and thus its magnitude for any item is tempered by the magnitude of the item's difficulty index.

A frequent problem encountered by "users" of item analyses is one of interpreting both indices, difficulty and discrimination, simultaneously and making a decision about the disposition of an item, either retaining or rejecting it for future use. All too frequently interpretations are confused when either or both indices depart, even slightly, from .50 for difficulty and positive or negative unity for discrimination. (It should be noted here that it is a popular misconception that the ideal difficulty index should be .50. For a comprehensive discussion of this point see Henrysson, 1971.)

The major objective of this paper is one of deriving a new index that

will facilitate the interpretation of item analyses. In this paper a new index,  $e$ , is presented as a function of difficulty and discrimination. Conceivably, the  $e$  index might facilitate as much if not more information than the simultaneous interpretations of difficulty and discrimination while at the same time it should be less confusing as it may be interpreted within a simple probability framework.

### *An Algebraic Rationale for Maximum Discrimination*

Assume that,  $a < .50$ , the item was difficult, then, less than one-half of the individuals responded positively to the item. It is possible, ideally, for all positive responses to have occurred in  $g_1$ , and all responses given by  $g_2$  would be negative responses leading to the inequality

$$r_1 + r_2 < n < w_1 + w_2.$$

Let the superscript \* denote maximum values. Then if  $(r_1 + r_2 < n)$  it may be assumed that all positive responses occurred in  $g_2$ . The maximum discrimination of an item,  $b^*$ , may now be defined algebraically as:

$$b^* = \frac{r_1 + r_2}{n}. \quad (3)$$

Inasmuch as  $(2n = N)$  the maximum discrimination of an item, given the difficulty  $(a \leq .50)$  of the item, may be written as

$$b^* = 2a; (a \leq .50). \quad (4)$$

When the observed difficulty of an item is less than or equal to .50, the maximum discrimination of the item is just two times the difficulty.

Assume that  $a > .50$ . Then less than one-half of the individuals responded negatively to the item. Implicitly

$$w_1 + w_2 < n < r_1 + r_2. \quad (5)$$

For purposes of determining a maximum discrimination index,  $b^*$ , it may be assumed that all negative responses occurred in  $g_2$ . Thus,  $r_1^*$  is assumed to be a maximum,  $r_1^* = n$ , and  $w_1^* = 0$ . The value for  $r_2^*$  may be computed as

$$r_2^* = (r_1 + r_2) - n. \quad (6)$$

Given the values for  $r_1^*$  and  $r_2^*$ , the maximum discrimination of an item with a difficulty greater than .50 may be defined specifically as

$$b^* = \frac{2n - (r_1 + r_2)}{n}. \quad (7)$$

In terms of observed difficulty

$$b^* = 2(1 - a); (a \geq .50). \quad (8)$$

When the observed difficulty of an item is greater than or equal to .50 the maximum discrimination of the item is equal to twice the proportion of negative responses.

When the difficulty of an item is less than .50 there are fewer positive responses than negative responses and the maximum discrimination index is a function of the negative responses. Alternatively, when the difficulty of an item is greater than .50 there are fewer negative responses than positive responses and the maximum discrimination index is a function of the negative responses.

As the difficulty index of an item deviates from .50, either above or below it, the maximum ceiling of the discrimination index is reduced from unity. For each metric unit of deviation from .50 for a difficulty index there is a two metric unit reduction from unity for the maximum ceiling of the associated discrimination index. Thus given a difficulty index of  $a$  its absolute deviation from .50,  $|d|$ , may be used to compute the ceiling or maximum possible discrimination index, in absolute value terms  $|b^*|$ , of the associated discrimination index.

$$|d| = |.50 - a| \quad (9)$$

and

$$|b^*| = 1.00 - 2|d|. \quad (10)$$

Logically the principle involved in computing maximum discrimination is presented by equations 9 and 10, however the pragmatics of the concept are obscured by the equations. Equations 4 and 8 represent a more reasonable set of equations for computing maximum discrimination.

### *A Cartesian Co-ordinate System Defined by Difficulty and Discrimination*

Geometrically maximum discrimination has a perfect curvilinear relationship to difficulty within the four quadrants of a two dimensional space. Within any one quadrant maximum discrimination is linearly related to difficulty. Because of an isomorphism between quadrants, the nonadjacent quadrants are reflections of each other and any pair of adjacent quadrants may be used to depict the relationship between difficulty and maximum discrimination.

In Figure 1, two adjacent quadrants of a Cartesian coordinate

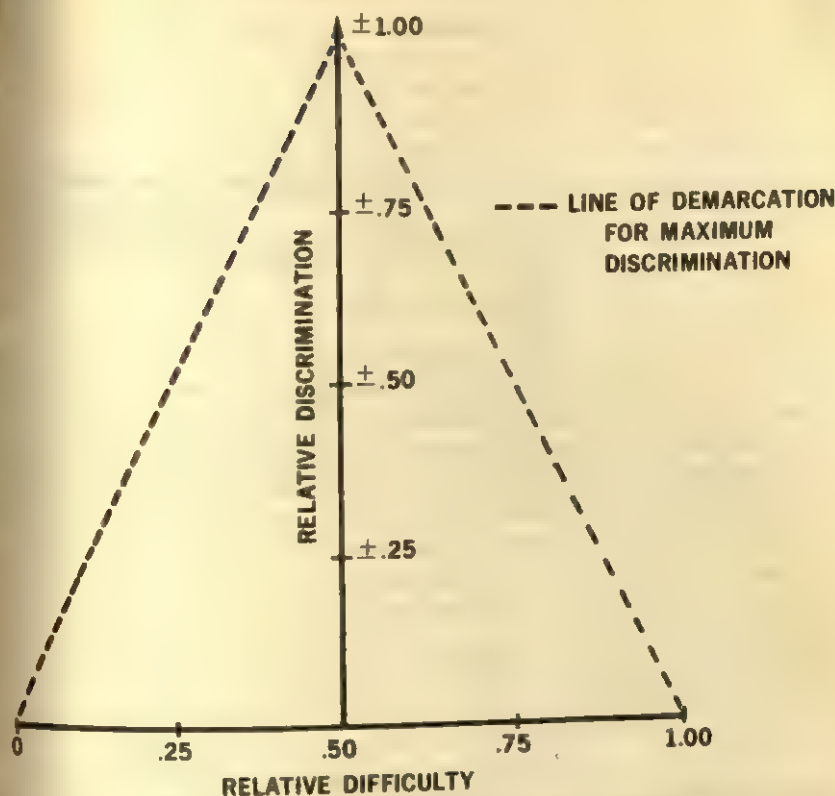


Figure 1. General Cartesian co-ordinate system defined by difficulty and discrimination.

system have been depicted. The abscissa of this system represents a difficulty continuum while the ordinate of the system represents a "dual signed" discrimination continuum, the values may be interpreted as either positive or negative.

The origin is denoted on the difficulty continuum as .50 so that any movement along the continuum will represent directed deviations from .50. The dashed line of demarcation within the "left quadrant" represents the line that is defined by any set of coordinates ( $a, b^*$ ) where ( $a \leq .50$ ) and  $b^*$  is a maximum discrimination value defined on the discrimination continuum. The dashed line in the "right quadrant" is the line that is defined by any set of coordinates ( $a, b^*$ ) where ( $a \geq .50$ ) and  $b^*$  is a maximum discrimination index, a value defined on the discrimination continuum.

In Figure 2 the terminus of an item vector,  $k$ , has been plotted with respect to the difficulty of the item,  $a$ , and the discrimination of the

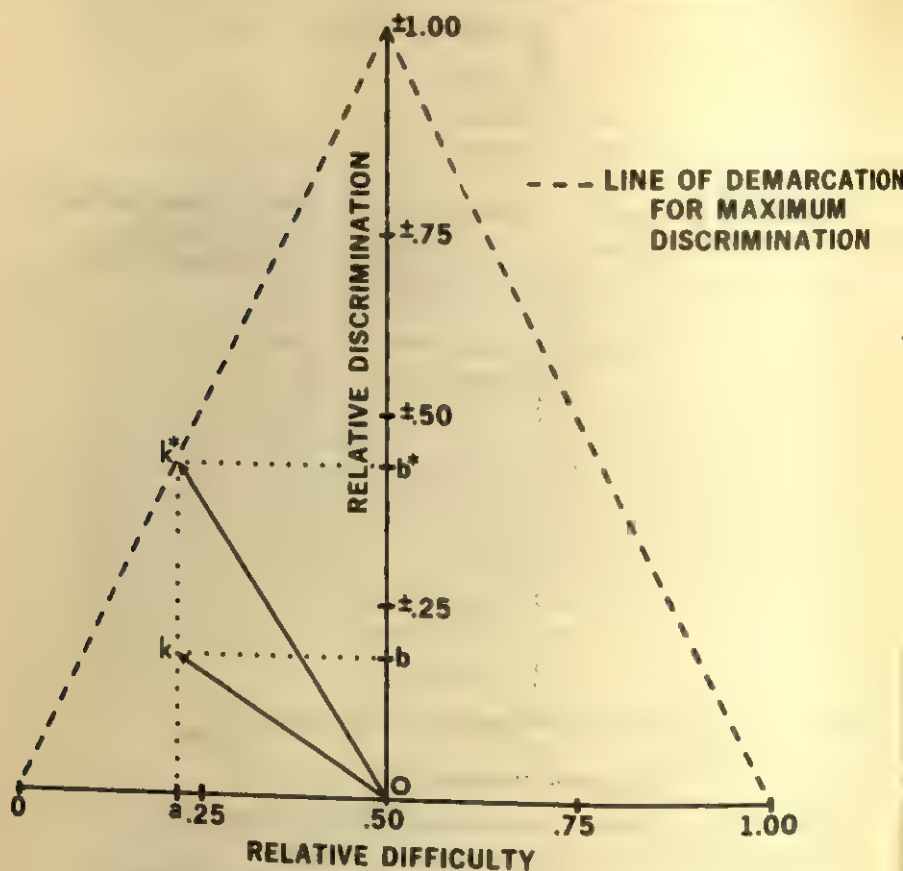


Figure 2. Item vector  $K$  defined by co-ordinates  $(a, b)$  and ideal item vector  $K^*$  defined by co-ordinates  $(a^*, b^*)$ .

item,  $b$ . A second item vector,  $k^*$ , has been plotted with respect to the difficulty of the item,  $a$ , and the maximum discrimination,  $b^*$ , of the item.

Assume the letter "o" is associated with the origin. There are two right triangles depicted in Figure 2,  $ako$  and  $ak^*o$ . Both triangles have the same base, the deviation of the difficulty index  $a$  from .50. The triangle defined by  $ak^*o$ , being the ideal triangle, will always be larger than triangle  $ako$ , the observed triangle. The ratio of the areas of the two triangles will indicate the size of  $ako$  relative to the maximum size it might have obtained. That is, the ratio of the two areas may be thought of as representing the efficiency of the item. The better an item functions, the more closely will the ratio of the two areas approach unity.



Technically, the area of the observed triangle is a function of the types of discriminations the item makes. In referring to Table 1 it may be observed that  $(r_1 + r_2)$  individuals are judged as being better than  $(w_1 + w_2)$  other individuals. However, the item functions as though  $[(r_1 + r_2)(w_1 + w_2)]$  dichotomous discriminations are made. The maximum number of discriminations which may be made for a given item is,  $(N/4)$ .

Consider the component parts of the equation

$$(r_1 + r_2)(w_1 + w_2) = r_1w_1 + r_1w_2 + r_2w_1 + r_2w_2; \quad (11)$$

then it is possible to consider the concept of "proper" and "improper" discriminations. Let the term proper discriminations refer to those point discriminations which are desirable in the sense that they result through a maximizing of the frequency of one particular response type, positive response, in one particular group, group one, while the other response type, negative response, is being maximized in the other group, group two. The term improper discrimination may be associated with those point discriminations which are not desirable in the sense that they occur as a result of undesirable response types occurring in both groups. For any item, the number of proper discriminations is characterized by  $(r_1w_2)$  and the improper discriminations by  $(r_2w_1)$ . Implicitly for easy items the negative responses should all be accrued by group two,  $w_2$ , and for difficult items the positive responses should all be accrued by group one,  $r_1$ .

A discrimination index in terms of point discriminations,  $\{b\}$ , is just the difference between proper and improper discriminations

$$\{b\} = (r_1w_2) - (r_2w_1) \quad (12)$$

and the relative discrimination index is given by

$$b = \frac{4\{b\}}{N^2}. \quad (13)$$

The maximum discrimination index, however, assumes no improper discriminations. Either  $r_2$  or  $w_1$  is assumed to be zero and either  $r_1$  or  $w_2$  is assumed to be  $n$ . Thus, maximum discrimination in terms of point discriminations,  $\{b^*\}$ , represents the maximum number of proper discriminations possible for a given  $N$  and difficulty index.

$$\{b^*\} = (r_1 + r_2)n; a \leq .50 \quad (14a)$$

or

$$\{b^*\} = (w_1 + w_2)n; a \geq .50 \quad (14b)$$

and the relative maximum discrimination index is given by

$$b^* = \frac{4\{b^*\}}{N^2}. \quad (15)$$

Given that the area of any triangle is equal to one-half the base multiplied by the altitude and given that both triangles in Figure 2 have the same base then the ratio of their areas is equivalent to the ratio of their altitudes. Alternatively, the ratio represents the number of observed proper discriminations less observed improper discriminations divided by the maximum possible number of proper discriminations. The ratio of  $\{b\}$  to  $\{b^*\}$  will range from zero to unity, assuming  $\{b\}$  is positive, and may be thought of, conceptually, as representing the "purity" of the discriminations made or the efficiency of the item. Let  $e$  represent a general efficiency index then:

$$e = \frac{\{b\}}{\{b^*\}} \quad (16)$$

and in modified form:

$$e = \frac{b}{b^*}. \quad (17)$$

The value for  $\{b^*\}$  will always be positive and  $\{b\}$  may be positive or negative, thus,  $e$  as defined by equations 16 and 17 may be positive or negative. When  $e$  is negative it is negative because more improper than proper discriminations were made. The terms proper and improper were somewhat arbitrarily assigned to two quantities on the assumption that more positive responses and, hence, fewer negative responses would always be made by group one relative to group two. For interpretations within the framework of proportions and areas, the sign of  $e$  may be neglected. The negative sign of  $e$  becomes meaningful only within the framework of probability.

Given the conditional magnitude of the difficulty index, the general  $e$  may be further specified as:

$$e_1 = \frac{b}{2a}; a \leq .50 \quad (18)$$

general efficiency for items having difficulty indices less than or equal to .50 and:

$$e_2 = \frac{b}{2(1-a)}; a \geq .50 \quad (19)$$

general efficiency for items having difficulty indices greater than or equal to .50.

Certain initial observations may be made with respect to  $e$  and proportion interpretations.

- (a) If the observed discrimination index of an item is zero then the efficiency of the item is zero.
- (b) For any level of difficulty, excluding zero and unity, it is theoretically possible for  $e$  to range from zero to unity assuming a positive discrimination index.
- (c) Efficiency is the ratio of observed proper discriminations less improper discriminations to the maximum possible number of proper discriminations for a given difficulty level and group size.
- (d) The general index  $e$  is indicative of how well an item has functioned relative to how well it might have functioned for a given  $N$  and specific difficulty level.

### *Probability Interpretations of Efficiency*

In the previous section it was noted that the general efficiency index could be subdivided into two indices, one for items having difficulties less than or equal to .50, henceforth efficiency of the first kind,  $e_1$  and one for items having difficulties greater than or equal to .50, henceforth efficiency of the second kind,  $e_2$ . The indices of efficiency may be further utilized to make probability interpretations with respect to positive responses and with respect to negative responses.

Equation 18 defining  $e_1$ , for items having difficulties less than or equal to .50, may be modified to define a computational equation for  $e_1$  regardless of item difficulty and sign of the discrimination index.

$$e_1 = \frac{r_1 - r_2}{r_1 + r_2} \quad (20)$$

Similarly, equation 19 defining  $e_2$ , for items having difficulties greater than or equal to .50, may be modified to define a computational equation for  $e_2$  regardless of item difficulty and sign of the discrimination index.

$$e_2 = \frac{w_2 - w_1}{w_1 + w_2} \quad (21)$$

Utilizing equations 20 and 21, it is possible to discuss conditional probabilities and note that quite unlike traditional U-L discrimination indices, efficiency considers two events which are mutually exclusive and exhaustive with respect to a given sample space.

Assume that a positive response has been made to an item. Given an individual making a positive response, the probability that the in-

dividual is a member of  $g_1$  is given by  $P(g_1 | r)$  while the probability that the individual is a member of  $g_2$  is given by  $P(g_2 | r)$ , where:

$$P(g_1 | r) = \frac{r_1}{r_1 + r_2} \quad (22)$$

and

$$P(g_2 | r) = \frac{r_2}{r_1 + r_2} \quad (23)$$

Then

$$e_1 = P(g_1 | r) - P(g_2 | r) \quad (24)$$

efficiency of the first kind is the difference between two conditional probabilities, where the probabilities are for group membership given a positive response.

Assume that a negative response has been made to an item. Given an individual making a negative response to an item, the probability that the individual is a member of  $g_1$  is given by  $P(g_1 | w)$  while the probability that the individual is a member of  $g_2$  is given by  $P(g_2 | w)$ , where:

$$P(g_1 | w) = \frac{w_1}{w_1 + w_2} \quad (25)$$

and

$$P(g_2 | w) = \frac{w_2}{w_1 + w_2} \quad (26)$$

Then

$$e_2 = P(g_2 | w) - P(g_1 | w) \quad (27)$$

efficiency of the second kind is the difference between two conditional probabilities, the probability of group membership given a negative response.

Efficiency of the first kind and efficiency of the second kind are both mutually exclusive and exhaustive with respect to sample space:

$$e_1 = P(g_1 | r) - P(g_2 | r); \quad (28)$$

$$1.0 = P(g_1 | r) + P(g_2 | r); \quad (29)$$

$$e_2 = P(g_2 | w) - P(g_1 | w); \quad (30)$$

$$1.0 = P(g_2 | w) + P(g_1 | w). \quad (31)$$

Also, it may be noted that  $e_1$  and  $e_2$  are proportional to each other. The ratio of  $e_2$  to  $e_1$  represents the odds in favor of a negative response

while the ratio of  $e_1$  to  $e_2$  represents the odds in favor of a positive response.

### *The Probability of Obtaining an Observed $e$ by Chance*

The general  $e$  index has been discussed with the framework of  $e_1$  and  $e_2$ . It was noted that for any given level of difficulty,  $e$  may range from zero to unity. Quite logically one would like to know the probability of obtaining an observed  $e$  for any given index of difficulty. Generically, what is a significant  $e$ ?

The model contingency table from which  $e$  is computed is unique within the framework of statistics. Theoretically,  $e$  is a measure of departure from independence in the contingency table. However, there is a different probability distribution for  $e$  associated with each uniquely different sample size,  $N$ , and each uniquely different difficulty level,  $a$ . In order to compute the probabilities associated with any given  $e$  it must be assumed that all four marginals of the contingency table are fixed. That is to say, the probabilities reported for any  $e$  are determined from the specific probability distribution of  $e$  associated with a given  $N$  and  $a$ .

Technically, in order to test the null hypothesis of independence, ( $e = 0$ ), it is necessary to compute the probability of obtaining the observed  $e$  and all possible  $e$  indices of a larger magnitude assuming constant marginals. Although Pearson's (1932) Chi square test might be used with this mode of a contingency table, it was not designed specifically for such use and in using it one would have to constantly keep in mind the consequences of its use with small sample sizes and also meet the assumptions of expected frequencies greater than five in the cells of the table.

A test designed specifically for the type of model contingency table associated with  $e$  is Fisher's (1935) exact test. Essentially, Fisher's test would indicate the exact probability of obtaining an observed  $e$  given a particular  $N$  and  $a$ . Furthermore, it could be used to compute the exact probabilities of each associated  $e$  greater than the observed  $e$ . In summing up all of these exact probabilities, one would have the probability of obtaining an  $e$  as large or larger than the one obtained for the given level of difficulty,  $a$ , and group size,  $N$ . Unfortunately, for any test having more than four or five items or fifteen or sixteen individuals, such an approach would be extremely time consuming.

However, it is possible to use a variation of Fisher's test, which is based upon the hypergeometric distribution, to establish the magnitudes of the  $e$  indices which would represent the extreme percentage of such indices for various difficulty levels and group sizes. Thus,



it is possible, so to speak, to establish "tables of significance" for the  $e$  index by computing probabilities associated with extreme values for  $e$  and work, computationally, toward the less extreme values.

Assuming Table 1 as a model, the total number of ways in which the table can be obtained, while maintaining fixed marginals is given by

$$C[N, (r_1 + r_2)]C[N, (n)] = \frac{(N!)^2}{(r_1 + r_2)! (w_1 + w_2)! (n!)^2} \quad (32)$$

which represents the product of the number of ways of taking  $(r_1 + r_2)$  responses from  $N$ , multiplied by the number of combinations of  $n$  individuals taken from the total  $N$  individuals. There are  $[N!/(r_1! r_2! w_1! w_2!)]$  ways of obtaining the cell frequencies in Table 1. Thus, the exact probability of obtaining the observed table frequencies and hence  $e$  may be computed as the ratio of the number of ways of obtaining the cell frequencies to the number of ways of taking  $(r_1 + r_2)$  from  $N$ , and then multiplied by the number of combinations of  $n$  individuals taken from the total  $N$  individuals which is:

$$P(e) = \frac{(r_1 + r_2)! (w_1 + w_2)! (n!)^2 n!}{r_1! r_2! w_1! w_2! N!} \quad (33)$$

In equation 33 it is important to note that the probability of either  $e_1$  or  $e_2$  is the same for any table inasmuch as the equation is associated with all four cells.

Assume that the difficulty of an item is less than .50, then it is possible to develop from equation 33 a computational algorithm for determining the significance of  $e$ -indices. Only the cell values  $r_1$ ,  $r_2$ ,  $w_1$  and  $w_2$  will change; thus, define as  $x$  that aspect of equation 33 that remains constant.

$$x = \frac{(r_1 + r_2)! (w_1 + w_2)! n! n!}{N!} \quad (34)$$

Assume as  $y_1$  the denominator of 33 excluding  $N!$ , that would be associated with the most extreme contingency table for the given  $N$  and  $a$ . The value of  $e$  associated with the most extreme contingency table would be unity. The probability of this value for the most extreme table may be determined from  $x$  and  $y_1$ . In this extreme case, the equation for  $y_1$  is given by

$$y_1 = (r_1 + r_2)! (0)! [n - (r_1 + r_2)]! (n)! \quad (35)$$

The exact probability of a table occurring with an efficiency index of unity given the particular  $N$  and  $a$  is  $x/y_1$ . (For continuity it is assumed that  $e = e_1$ ,  $a \leq .50$ , however, equations 35-40 may be used for  $e = e_2$

by simply substituting  $w_1$  for  $r_1$  and  $w_2$  for  $r_2$ .) Compute the  $y$ -value for the subsequently more independent tables following the computational procedure

$$\begin{aligned} y_2 &= (r_1 + r_2 - 1)! (1)! [n - (r_1 + r_2 - 1)]! (n - 1)! \\ y_3 &= (r_1 + r_2 - 2)! (2)! [n - (r_1 + r_2 - 2)]! (n - 2)! \\ &\vdots \\ y_j &= [r_1 + r_2 - (j - 1)]! (j - 1)! [n - (r_1 + r_2 - (j - 1))]! (n - (j - 1))! \end{aligned} \quad (36)$$

In a special sense, when using Fisher's (1935) exact test, one first computes  $y_j$  and then  $y_{j-1}$  and so on to  $y_1$ . Essentially the exact probability,  $\alpha$ , of obtaining the cells of an observed contingency table or some worse departure from independence may be computed by

$$\alpha = \frac{x}{\sum_{i=1}^j y_i} \quad (37)$$

Implicit in Fisher's test is the assumption of finite sample size. To establish an  $\alpha$ -level and then attempt to determine the associated  $e$ -value is tantamount to assuming an infinite sample size. The assumption of finite sample size is a restriction that necessarily exists because equations 33, 34, 35, and 36 use discrete numbers. Thus, for any sample of "modest" size, say  $N = 60$ , one is faced with the task of computing extremely large factorials.

When two of the fixed marginals are identical, as they are in Table 1, the computation of probability levels and associated  $e$ -indices is greatly simplified. There are two rather compelling properties associated with such a contingency table. First, the hypergeometric probability distribution associated with such a table will be symmetric. Secondly, the most extreme departure from independence is immediately known, at least one of the four cells will be zero and the efficiency index will be unity.

Within the framework of this paper one may think of Fisher's exact probability test as determining the exact probability of obtaining an  $e$ -index as large or larger than the one observed. However, through an inexact procedure it is possible to establish a critical probability, say  $\alpha$ , and then compute the discrete  $e$ -index whose exact probability,  $\alpha'$ , is the best estimate of  $\alpha$  given the restriction that, ( $\alpha' \leq \alpha$ ), the observed probability level may not be larger than  $\alpha$ . An approach such as this would indirectly facilitate the use of much larger, greater than 60, but still finite sample sizes.

Assume ( $\alpha$ ); then  $x$  by equation 34, and  $y_1$  by equation 35. If  $x/y_1 < \alpha$  then compute  $y_2$  and so on following equation 36 until for the  $j$ th and  $(j+1)$  values for  $y$  the following inequality occurs.

$$\frac{x}{\sum_{i=1}^j y_i} \leq \alpha \leq \frac{x}{\sum_{i=1}^{j+1} y_i} \quad (38)$$

The above inequality implies that the probability of the  $e$ -index associated with the cells of the  $j$ th contingency table is less than or equal to  $\alpha$  while the probability of the  $e$ -index associated with the cells of the  $(j+1)$ th contingency table is greater than  $\alpha$ . The following two equations may be used to compute  $\alpha'$  and  $e'$  respectively:

$$\alpha' = \frac{x}{\sum_{i=1}^j y_i}; \quad (39)$$

$$e' = 1 + \frac{2(1 - j)}{r_1 + r_2}, \quad (40)$$

where one may correctly assume that the probability, not necessarily exact, of obtaining an  $e$  index as large as or larger than  $e'$  given a particular level of difficulty is equal to or less than  $\alpha$ . It is prudent to realize that the discussion here applies to only one tail of the symmetric distribution. If one considers a difficulty greater than .50 it is only necessary to substitute  $w_1$  for  $r_1$  and  $w_2$  for  $r_2$  in equations 35-40. Alternatively, one tail of the distribution is associated with  $e_1$  while the other tail is associated with  $e_2$ . One may consider just  $e_1$  and use the odds ratio to convert an observed  $e_1$  to its associated  $e_2$  value and use precisely the same probability interpretations for  $e_2$  as those associated with  $e_1$ . For a difficulty index of .50, it makes no difference how the probabilities are computed inasmuch as ( $e_1 = e_2$ ). (A table of .05 critical values for  $e'$  across all levels of difficulty for samples ranging in size up to 100 is available from the author.)

### *Total Test Efficiency*

This paper has been concerned primarily with the computation of the  $e$ -index at an item level. This does not, however, preclude its use as a total test statistic. Just as one can talk of total test difficulty, so, also, can one talk of total test efficiency. In this section, the equations for computing total test efficiency will be discussed. No attempt will be

made to interpret the total test efficiency index other than the cursory definition that follows from it computationally.

Assume some test composed of  $j$  items. Then  $N_j'$  individuals will respond to item  $j$ . The number of individuals in either  $g_1$  or  $g_2$  will be denoted as  $n_j$  for the  $j$ th item and ( $2n_j = N_j'$ ). The difficulty and discrimination indices for the  $j$ th item may be denoted as  $a_j$  and  $d_j$  respectively. The total number of possible discriminations that may be made by the  $j$ th item is  $N_j^2/4$ . The absolute maximum frequency of proper discriminations that is possible for the  $j$ th item,  $b_j^*$ , is given as:

$$\{b_j^*\} = N_j a_j. \quad (41)$$

The frequency of proper discriminations less improper discriminations,  $\{b_j\}$ , for the  $j$ th item is given as:

$$\{b_j\} = n_j b_j. \quad (42)$$

Inasmuch as efficiency is defined, at an item level, as the ratio of observed proper discriminations less improper discriminations to the maximum possible number of proper discriminations, for a given difficulty index and group size, assume a similar definition for total test efficiency. Let total test efficiency be represented by the ratio of total observed proper discriminations less improper discriminations to the maximum possible number of proper discriminations for the total test. Let  $E_1$  represent total test efficiency, assuming that the difficulty of all items is less than .50, then

$$E_1 = \frac{\sum_{i=1}^j n_i b_i}{2 \sum_{i=1}^j n_i a_i}. \quad (43)$$

Such an index is indicative of the proportion of "quality" discriminations made by a total test given that all items have a difficulty less than .50. For items having a difficulty index greater than .50 equation 43 is modified to determine  $E_2$  as

$$E_2 = \frac{\sum_{i=1}^j n_i b_i}{2 \sum_{i=1}^j n_i (1 - a_i)}. \quad (44)$$

If the  $j$  items form a mastery test, then either equation 43 or 44 could be used to compute total test efficiency. If the  $j$  items form an achievement test, assume that  $s$  of the items have difficulty indices less than .50 and assume that  $k$  of the items have difficulty indices greater than or

equal to .50, then if the items are grouped dichotomously according to those items having difficulty indices less than .50 and those equal to or greater than .50 the general efficiency index,  $E$ , of the achievement test is given by:

$$E = \frac{\sum_{i=1}^i n_i b_i}{2 \left[ \sum_{i=1}^i n_i a_i + \sum_{i=1}^k n_i (1 - a_i) \right]} \quad (45)$$

Experience with this index is still growing and, therefore, its application to total tests is only of theoretical interest at this time. However, the systematic study of  $E$ ,  $E_1$  and  $E_2$  might be informative. To wit, either  $E_1$  or  $E_2$  might serve as indicators of discriminatory of difficulty homogeneity or perhaps as some sort of index of internal consistency for a mastery test. The general index,  $E$ , might serve as an index of internal consistency for an achievement test.

#### *A Pragmatic Scheme for the Use of $e$*

It is possible to set up subjective criteria for determining easy, moderate, and hard items as well as nonefficient, efficient, and ideally efficient items. Let the following inequalities, based upon efficiency,  $e$ , and difficulty,  $a$ , serve as operational definitions of the above terms.

$0.00 \leq e \leq 0.50$	nonefficient item
$0.50 \leq e \leq 0.80$	efficient item
$0.80 \leq e \leq 1.00$	ideally efficient item
$0.75 \leq a \leq 1.00$	easy item
$0.25 \leq a \leq 0.75$	moderate item
$0.00 \leq a \leq 0.25$	hard item

Because exact intervals have not been made these definitions are, theoretically, not mutually exclusive but for practical purposes they may be thought of as mutually exclusive and exhaustive with respect to difficulty and efficiency. Utilizing these definitions, one may categorize all items of an instrument into one of nine subjective item types, e.g., nonefficient easy items.

Based upon this categorization it is possible to construct within the framework of Cartesian coordinates a chart for determining item quality. In Figure 3 such a chart has been constructed. Given the difficulty and discrimination of an item as coordinates, it is possible to



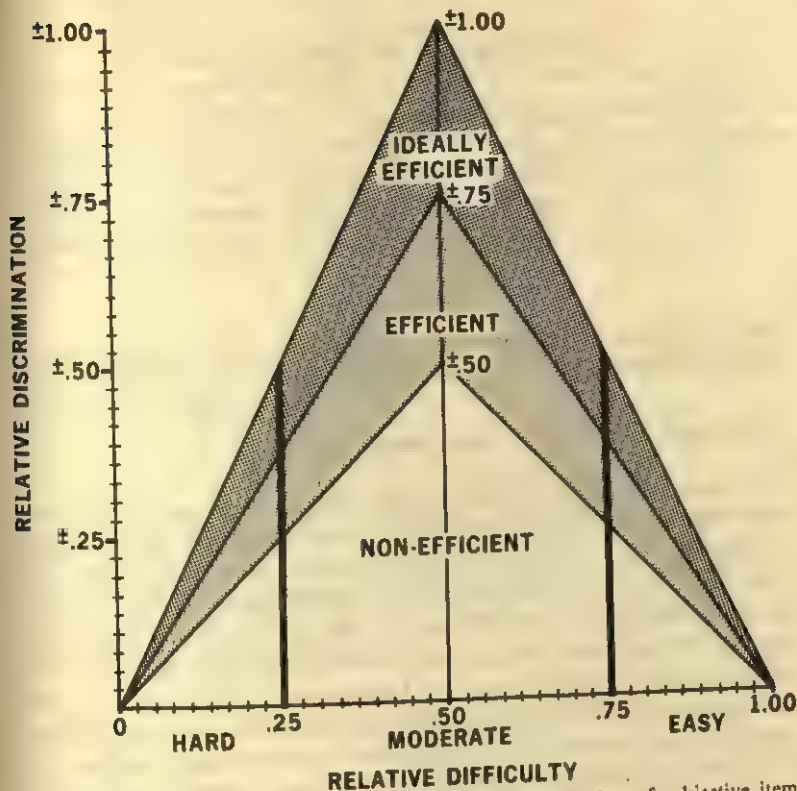


Figure 3. Iconic representation of contingency categorization of subjective item types.

locate the associated item point. If the point falls within the working area, then the item is a working item. Additionally, it is possible to determine the relative difficulty of the working item. Although this chart was constructed as a function of the *e*-index, its use actually precludes the computation of such an index.

Assuming Figure 3 to be triangular probability distribution, it is possible to briefly discuss the chance probability of obtaining different item types. Note in Figure 3 that there are nine different item types. The probability of any item types occurring by chance may be determined as the ratio of the surface area associated with each item type to the total surface area of the probability distribution. The probabilities for the nine subjective item types are reported in Table 2.

As previously noted, the concept of item types is subjective as are the arbitrary numerical criteria for defining them. It is possible to con-

TABLE 2  
*Chance Probabilities of Subjective Item Types*

Difficulty Levels	Nonefficient	Efficient	Ideally Efficient	Total
Easy	.0625	.0375	.0250	.1250
Moderate	.3750	.2250	.1500	.7500
Hard	.0625	.0375	.0250	.1250
Total	.500	.300	.200	1.0000

struct a more detailed chart than Figure 3, as well as a more detailed table of chance probabilities to go with it.

### *The e-index and Certain other Indices*

If certain assumptions pertaining to substantial sample size are met, the *e*-index may be converted to a Pearson  $\chi^2$  statistic. The problem of sample size is closely related to the problem of small sample size expectancies in the  $2 \times 2$  table. Failure to meet Pearson's (1932, 1916) assumption of sample size will result in a statistic following some hypergeometric distribution rather than a chi-square distribution. Consider the marginals of Table 1, one set of marginals is always fixed, within a proportional framework, at  $n/N$ . The other set of marginals, within a proportional framework, vary between zero and unity defining the item difficulty and its complement. Let five be the lower bound for expected cell frequency, then  $5/N$  would represent the expected cell proportion. Assuming independence an expected cell proportion may be computed as the product of the two marginals associated with the cell. The minimum difficulty of an item for a valid use of a Pearson chi square test is then given by

$$a \geq \frac{10}{N} \quad (46)$$

Assuming an item has a difficulty greater than or equal to  $10/N$  and less than or equal to  $(1 - 10/N)$  a Pearson chi-square statistic, actually a test of homogeneity, may be applied to the item's contingency table. Specifically, the conditional inequality

$$\frac{10}{N} \leq a \leq \left(1 - \frac{10}{N}\right) \quad (47)$$

must be met, assuming  $N$  is sizable. Assuming these conditions have been met then  $\chi^2$ , the Pearson chi square statistic may be computed directly from either  $e_1$  or  $e_2$ , with 1 degree of freedom.

$$x^2 = (Ne_1^2) \left( \frac{a}{1-a} \right) = (Ne_2^2) \left( \frac{1-a}{a} \right) \quad (48)$$

A second index which is closely related to efficiency, at least algebraically, is the phi index. More specifically, the ratio of the phi index to its maximum possible value phi-max, as defined by Cureton (1959) is algebraically identical to the efficiency index.

Given the equation defining phi as a function of  $\chi^2$ :

$$\phi = \left[ \frac{x^2}{N} \right]^{1/2} \quad (49)$$

it is possible by simple substitution in equation 48 to define phi as a function of efficiency.

$$\phi = e_1 \left( \frac{a}{1-a} \right)^{1/2} = e_2 \left( \frac{1-a}{a} \right)^{1/2} \quad (50)$$

Following Cureton (1959), it is also possible to define the maximum value for  $\phi$  from Table 1 within the context of difficulty. Just as a distinction between  $e_1$  and  $e_2$  occurred as a function of the difficulty being greater than or less than .50 so also must this distinction be made for the maximum phi coefficient.

$$\phi_{\max} = \left( \frac{a}{1-a} \right)^{1/2} ; a \leq .50; \quad (51a)$$

$$\phi_{\max} = \left( \frac{1-a}{a} \right)^{1/2} ; a \geq .50. \quad (51b)$$

In forming the ratio of phi to its maximum value, two equations result as a function of difficulty. The first equation results as a function of ( $a \leq .50$ ) and is

$$\frac{\phi}{\phi_{\max}} = e_1 \quad (52)$$

just a redefinition of  $e_1$  in terms of phi and its maximum. The second ratio being a function of ( $a \geq .50$ ) is just

$$\frac{\phi}{\phi_{\max}} = e_2 \quad (53)$$

a redefinition of  $e_2$  in terms of phi and its maximum.

It is important to note here that the computation of  $\phi_{\max}$  must follow 51 if the item difficulty is less than or equal to .50 and the computation of  $\phi_{\max}$  must follow 51 if the item difficulty is greater than or equal to .50.

Although there is an apparent close algebraic relationship between

efficiency and the ratio  $(\phi/\phi_{\max})$ , the concept of general efficiency, the use of both  $e_1$  and  $e_2$ , would appear to be more meaningful than the ratio  $(\phi/\phi_{\max})$ . We also find ourselves hard pressed to interpret general efficiency within the framework of a Pearsonian correlation. Finally, based upon Carroll's (1961) discussion of  $(\phi/\phi_{\max})$  it would seem most prudent to caution against forcing efficiency into the framework of the phi coefficient.

### Conclusion

This paper has presented the basic aspects of the efficiency index. Whether or not the index will replace or supplement the traditional discrimination index remains to be seen. However, it would seem that the general efficiency index has many more statistically compelling properties than the traditional discrimination index.

### REFERENCES

- Carroll, J. B. The nature of the data, or how to choose a correlation coefficient. *Psychometrika*, 1961, 26, 347-372.
- Cureton, E. E. The upper and lower twenty-seven per-cent rule. *Psychometrika*, 1959, 22, 293-296.
- Fisher, R. A. The logic of inductive inference. *Journal of Royal Statistical Society*, 1935, 98, 39-54.
- Henryssen, S. Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.), *Educational measurement*. (2nd ed.) Washington, D. C.: American Council on Education, 1971. pp. 130-159.
- Johnson, A. P. Notes on a suggested index of item validity: The U-L index. *Journal of Educational Psychology*, 1951, 42, 499-504.
- Kelley, T. L. The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, 1939, 40, 17-19.
- Pearson, K. On the general theory of multiple contingency with special reference to partial contingency. *Biometrika*, 1916, 11, 145-158.
- Pearson, K. Experimental discussion of the  $(\chi^2, P)$  test for goodness of fit. *Biometrika*, 1932, 24, 351-381.

## FACTOR STRUCTURE OF THE MCCARTHY SCALES AT FIVE AGE LEVELS BETWEEN 2½ AND 8½<sup>1</sup>

ALAN S. KAUFMAN<sup>2</sup>

The Psychological Corporation

The McCarthy Scales of Children's Abilities (MSCA) were factor analyzed at five age levels: 2½, 3-3½, 4-4½, 5-5½, and 6½-7½-8½. The standardization sample ( $N = 1032$ ) provided the source of data. Varimax rotated factors akin to four of the six MSCA Scales—General Cognitive, Verbal, Memory, and Motor—appeared at age 2½, and tended to appear at all older age levels. Factors akin to the Perceptual-Performance and Quantitative Scales emerged at ages 3-3½ and 5-5½, respectively. The overall findings were interpreted from a developmental perspective, and the data were shown to offer evidence for the construct validity of the MSCA.

THE McCarthy Scales of Children's Abilities (MSCA) comprise 18 short mental and motor tests which have been grouped into six scales: Verbal, Perceptual-Performance, Quantitative, General Cognitive, Memory, and Motor. McCarthy (1972) selected the scales primarily on the basis of functional and intuitive considerations, although the results of preliminary factor analyses of parts of the MSCA standardization data were also considered. These analyses of the standardization edition of the MSCA, though exploratory in nature, served a number of other useful functions. They suggested that the battery (1) has a strong underlying structure, as evidenced by the consistency of the factor patterns for each set of data when several different techniques of factor analysis were applied; (2) had a somewhat similar structure at three different age levels; and (3) measures some abilities

<sup>1</sup> I am particularly grateful to Dr. Dorothea McCarthy for her helpful suggestions regarding the implications of the results of the study.

<sup>2</sup> Reprints may be obtained from Alan S. Kaufman, Department of Educational Psychology, College of Education, University of Georgia, Athens, Georgia 30602.



which are similar to the abilities assessed by conventional intelligence tests, and others which seem to add uniqueness to the field of children's testing (Kaufman and Hollenbeck, 1973).

The purpose of the present study was to analyze the final version of the MSCA (which is somewhat shorter than the standardization edition) to provide a more definitive picture of the MSCA's factor structure. The availability of data for the entire standardization sample made it possible to have a relatively large ratio of the number of subjects to the number of variables in each of the present analyses. Such ratios help insure the stability of the resulting factor patterns, and therefore permit more meaningful comparisons of the factor structure at different age levels.

In addition to gaining this developmental perspective of the MSCA, and relating the results to existing theory and research, a second major purpose of the study was to evaluate the construct validity of the McCarthy Scales; as Anastasi (1968, pp. 114-120) indicates, factor analysis is one of the acceptable techniques for providing evidence of a test's construct validity. Since McCarthy's (1972, p. 2) main goal in structuring the scales was to develop a clinically useful instrument (rather than a factorially pure test), the construct validity of the MSCA certainly does not rest on data obtained from factor analysis. Nevertheless, a close correspondence between the factor structure and the chosen scales will enhance the instrument's validity, and should make the scores more meaningful to the clinician.

## *Method*

### *Instrument*

As indicated, the McCarthy Scales include a number of tests which have been grouped into six scales. Figure 1 provides a schematic illustration of the content of each scale and the interrelationship that exists among the scales.

The component tests are described in detail in the manual (McCarthy, 1972), and are merely summarized here: (1) *Block Building*—copying structures made out of cubes, (2) *Puzzle Solving*—putting together cut-up pictures, (3) *Pictorial Memory*—recalling pictures exposed briefly, (4) *Word Knowledge*—Picture Vocabulary (Part I) and Oral Vocabulary (Part II), (5) *Number Questions*—number facts and oral problems, (6) *Tapping Sequence*—repeating sequences tapped on a xylophone, (7) *Verbal Memory*—repeating words and sentences (Part I) and retelling a story (Part II), (8) *Right-Left Orientation*—knowing right vs. left, (9) *Leg Coordination*—motor skills such as

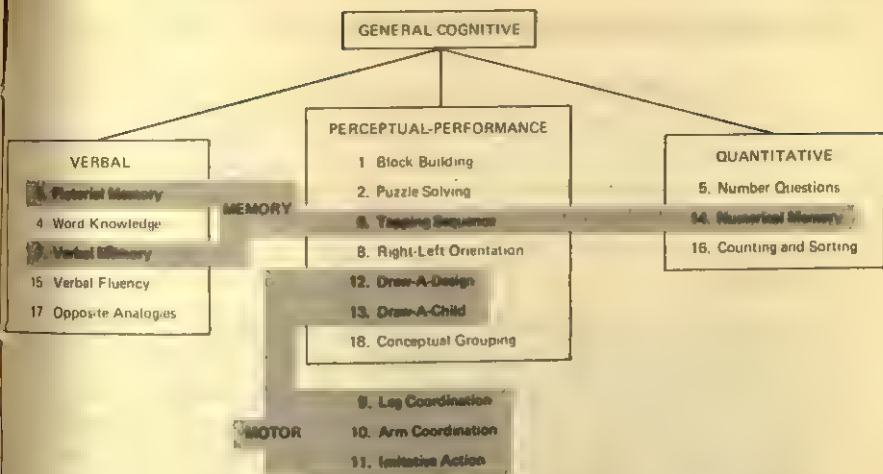


Figure 1. The grouping of the 18 MSCA tests into six scales. The tests in the Verbal (V), Perceptual-Performance (P), and Quantitative (Q) Scales are combined to form the General Cognitive Scale. Each of the Memory tests is also included on either the V, P, or Q Scale, and hence on the General Cognitive Scale. The Motor Scale contains three non-cognitive tests which belong exclusively to it, and two tests which are shared with other scales. (Thanks are due Mrs. Fay B. Krawchick for designing the figure.)

walking backwards, (10) *Arm Coordination*—bouncing a ball (Part I), catching a beanbag (Part II), and throwing a beanbag at a target (Part III), (11) *Imitative Action*—simple motor skills such as clasping hands, (12) *Draw-A-Design*—copying designs, (13) *Draw-A-Child*—drawing a child of the same sex, (14) *Numerical Memory*—repeating digits forwards (Part I) and backwards (Part II), (15) *Verbal Fluency*—naming as many “things” in several categories as possible in 20 seconds, (16) *Counting and Sorting*—simple number concepts, (17) *Opposite Analogies*—providing opposites to complete analogies, (18) *Conceptual Grouping*—logical classification.

### Subjects

The standardization sample of the MSCA, which includes 100 to 106 children at each of 10 age levels between 2½ and 8½ (Total  $N = 1032$ ), provided the source of data. At each age level there were an equal number of boys and girls and a proportional representation of whites and nonwhites in accordance with 1970 Census data. A detailed description of the sample and the stratification variables appears in the manual (McCarthy, 1972). For the present analyses, the sample was

divided into the five groups shown below:

Age Levels	<i>N</i>
2½	102
3-3½	204
4-4½	206
5-5½	206
6½-7½-8½	314

### *Procedure*

The tests constituting the MSCA were the variables studied. Four of the tests include 2 or 3 parts and, in general, these parts were treated as separate variables in the present factor analyses. Whenever a test or part of a test produced virtually no variation in test scores for a particular age group (e.g., nearly every 2½ year-old scored 0 on Draw-A-Child), then that variable was excluded from the analysis for the age group in question. Word Knowledge was an exception to both of these rules. For this test, the total of Parts I and II was entered in each matrix since Part I produced little or no variation in test scores at all but the youngest age levels.

The results of the exploratory factor analyses of the standardization edition of the MSCA (Kaufman and Hollenbeck, 1973) helped to guide the procedure for the present investigation. Since the several different factor analytic techniques used for each set of data in the previous study gave consistent results, the use of only one technique seemed sufficient for the present analysis. In addition, the use of objective methods for determining the number of factors led to the extraction of a few apparently trivial factors. Therefore, for the present analyses, the solutions which made the most psychological sense were selected.

For the analyses, correlation matrices were obtained for the five age groups (which will be referred to as ages 2½, 3, 4, 5, and 6+). Then, each matrix was subjected to principal factor analysis, with squared multiple correlations in the diagonals, followed by varimax rotation of 3-, 4-, 5-, and 6-factor solutions.

### *Results*

#### *Factor Structure at the Five Levels*

The following solutions seemed to make the most psychological sense and were, therefore, selected: 4-factor at age 2½, 5-factor at age 5, and 6-factor at ages 3, 4, and 6+. Loadings of .25 and above

were interpreted as meaningful and only these loadings are included in the tables.

*Age 2½.* (See Table 1.) As is shown, the four factors were called *Verbal*, *Motor*, *General Cognitive*, and *Memory*, respectively. Factor II is a clear motor factor as each of the three gross motor tests had loadings of about .60. The high loadings of Pictorial Memory and Verbal Memory on Factor IV suggested that much of the ability involved is memory. Factors I and III are both cognitive in nature and each accounted for 27% of the common factor variance. Factor I was called *Verbal* because the tasks with the highest loadings involve verbal ability, and all tasks with loadings of .25 or better (except for Arm Coordination), require either verbalization by the child or comprehension of verbal questions and commands. Factor III was called *General Cognitive* because all of the tasks with high loadings require conceptualization, whether verbal, nonverbal, or quantitative. The only tasks which did *not* have meaningful loadings were the ones assessing either gross motor coordination or simple rote memory.

*Age 3.* (See Table 2.) Factor I was called *General Cognitive*, as most of the cognitive tasks in the battery had substantial loadings. Although

TABLE 1  
*Varimax Rotated Factor Matrix of the MSCA Tests at Age 2½*

Test	Factor			
	I Verbal	II Motor	III General Cognitive	IV Memory
Block Building		.30	.42	.45
Puzzle Solving			.48	.35
Pictorial Memory				.62
Word Knowledge I + II	.34	.25	.49	.29
Number Questions	.41	.31	.30	.25
Tapping Sequence		.44		.28
Verbal Memory I	.38			.54
Leg Coordination		.61		
Arm Coordination	.43	.59		
I + II + III				
Imitative Action		.58		
Draw-A-Design		.32	.37	
Numerical Memory I	.57		.38	
Verbal Fluency	.58	.41	.62	
Counting and Sorting			.38	
Opposite Analogies	.65		.57	
Conceptual Grouping	.40			.30
% of Com. Fact. Var.	27	26	27	20

Note.—Only loadings of .25 and above are included.

TABLE 2  
*Varimax Rotated Factor Matrix of the MSCA Tests at Ages 3-3½*  
 (N = 204)

Test	Factor					
	I General Cognitive	II Motor	III Verbal	IV Memory	V Drawing	VI Perceptual- Performance
Block Building	.25	.26				.42
Puzzle Solving			.32	.48		.36
Pictorial Memory			.64	.30		
Word Knowledge I + II	.63			.38		
Number Questions	.75					
Tapping Sequence	.32			.33		
Verbal Memory I	.26		.40	.54		
Verbal Memory II	.26			.68		
Leg Coordination	.37					
Arm Coordination I		.61				
Arm Coordination II		.65				.30
Arm Coordination III	.30	.52				
Imitative Action	.30	.28				.26
Draw-A-Design		.25	.43		.43	
Draw-A-Child	.39				.55	
Numerical Memory I	.41		.27	.27	.32	
Verbal Fluency			.63			
Counting and Sorting	.37		.54			
Opposite Analogies	.50		.44	.30		
Conceptual Grouping	.50		.45			.29
% of Com. Fact. Var.	26	16	23	18	9	8

Note.—Only loadings of .25 and above are included.

a few of the gross motor tasks had loadings of .25 or above, it is clear that the very high loadings belonged to conceptual tasks such as Number Questions and Word Knowledge. Factor II was interpreted as a *Motor* factor, and Factor III as a *Verbal* factor. Factor IV was called *Memory*, as Verbal Memory I and II had the highest loadings and all tests of short-term memory had meaningful loadings.

Factors V and VI each had high loadings by some of the tasks involving perceptual-motor coordination. Factor V is fairly specific in content and was termed *Drawing* as the 2 drawing tests had the highest loadings. Factor VI was labeled *Perceptual-Performance* since it is more varied with the nonverbal Block Building, Puzzle Solving, and Conceptual Grouping having 3 of the 4 highest loadings.

*Age 4.* (See Table 3.) As is evident, no general factor was found in this analysis. Of the six factors, three were easily identifiable: Factor I was called *Drawing* because of the high loadings by the drawing tests; Factor II is a clear *Motor* factor; and Factor V was called *Perceptual-Performance* due to the high loadings by the cognitive tasks requiring



TABLE 3  
*Varimax Rotated Factor Matrix of the MSCA Tests at Ages 4-4½*  
 (N = 206)

Test	Factor					
	I	II	III	IV	V	VI
	Drawing	Motor	Memory	Verbal	Perceptual- Performance	Semantic Memory
Block Building					.44	
Puzzle Solving	.33		.26		.38	
Pictorial Memory						.49
Word Knowledge			.49	.40		.27
Number Questions				.53		.26
Tapping Sequence	.28		.39		.50	
Verbal Memory I			.67			.30
Verbal Memory II			.40		.26	.38
Leg Coordination		.32				
Arm Coordination I		.50				
Arm Coordination II		.59				
Arm Coordination III		.39				
Imitative Action		.25			.28	
Draw-A-Design	.51			.25	.34	
Draw-A-Child	.56			.25	.26	
Numerical Memory I	.26		.48	.30		
Verbal Fluency			.45	.39	.31	
Counting and Sorting				.31	.57	
Opposite Analogies				.59		
Conceptual Grouping			.32	.55	.34	
% of Com. Fact. Var.	14	13	22	21	20	10

Note.—Only loadings of .25 or above are included.

no verbal responses by the child. (The .57 loading by Counting and Sorting on Factor V is not inconsistent with the interpretation because this number task requires virtually no verbalization.)

The remaining three factors all seem to deal with verbal ability, with two of them also involving memory. Factor IV was termed *Verbal* because of its close similarity with the Verbal factors identified at ages 2½ and 3. Factor III was called *Memory* since most of the tasks with high loadings involve short-term memory (Verbal Memory I & II, Numerical Memory, Tapping Sequence). Finally, Factor VI was labeled *Semantic Memory* since the three variables with the highest loadings are all memory tasks which have a semantic content.

Age 5. (See Table 4.) Factor I was given the name *General Cognitive/Verbal*. This factor, which accounts for 37% of the common factor variance, is certainly a general factor; however, since five of the six tasks with loadings greater than .50 require verbal ability, the dual name was assigned.

Factors II, III, and V seem to be clear *Motor*, *Perceptual-*

TABLE 4  
*Varimax Rotated Factor Matrix of the MSCA Tests at Ages 5-5½*  
 (N = 206)

Test	I Gen. Cog./ Verbal	II Motor	Factor III Perceptual- Performance	IV Memory	V Quantitative
Block Building			.37		
Puzzle Solving	.25		.50		
Pictorial Memory				.42	
Word Knowledge I + II	.54		.28		
Number Questions	.40		.27		.46
Tapping Sequence				.25	.37
Verbal Memory I	.64				
Verbal Memory II	.54				
Right-Left Orientation					
Leg Coordination	.34	.32			
Arm Coordination I		.53			
Arm Coordination II		.57			
Arm Coordination III		.44			
Imitative Action				.26	
Draw-A-Design			.56		.26
Draw-A-Child	.30		.57		
Numerical Memory I	.56				.35
Numerical Memory II	.27	.28	.34		.41
Verbal Fluency	.51		.33		
Counting and Sorting	.34		.43		.47
Opposite Analogies	.58				
Conceptual Grouping	.49		.42		.25
% of Com. Fact. Var.	37	14	25	8	16

Note.—Only loadings of .25 or above are included.

*Performance*, and *Quantitative* factors, respectively. Factor IV was labeled *Memory*, although it is more specific than the memory factors found at the younger age levels, and seems to involve visual memory. The large memory factor found at age 5 in the analyses of the standardization edition of the MSCA was highlighted by the substantial loadings of Numerical Memory I, Verbal Memory I, and various tests of verbal ability (Kaufman and Hollenbeck, 1973). That factor seems to have merged with the General Cognitive factor in the present analysis.

*Age 6+.* (See Table 5.) As with the analysis at age 5, Factor I was given the dual name *General Cognitive/Verbal*. Factors II, III, IV, and V are clearly identifiable as *Motor*, *Perceptual-Performance*, *Memory*, and *Quantitative*, respectively. Factor V was called *Reasoning* because the tasks with meaningful loadings tend to require the child to conceptualize at a higher level than most other tasks in the battery.

TABLE 5  
*Varimax Rotated Factor Matrix of the MSCA Tests at Ages 6½-7½-8½*  
 (N = 314)

Test	Factor					
	I Gen.Cog./ Verbal	II Motor	III Perceptual Performance	IV Memory	V Reasoning	VI Quantitative
Puzzle Solving	.26	.26	.45		.39	
Pictorial Memory	.46					
Word Knowledge I + II	.66		.37		.28	
Number Questions	.36	.28	.38	.32	.31	.29
Tapping Sequence		.30	.29			
Verbal Memory I	.56			.31		
Verbal Memory II	.58					
Right-Left Orientation						
Leg Coordination		.27				
Arm Coordination I		.57				
Arm Coordination II		.60				
Arm Coordination III		.49				
Draw-A-Design			.59			
Draw-A-Child	.36		.49			
Numerical Memory I				.50		
Numerical Memory II	.25		.48			.29
Verbal Fluency	.55		.28			
Counting and Sorting			.26			.42
Opposite Analogies	.44		.37		.27	.29
Conceptual Grouping					.45	
% of Com. Fact. Var.	30	18	23	9	10	10

Note.—Only loadings of .25 and above are included.

### Discussion

Factor analytic studies of the abilities of school-age children have appeared frequently in the literature, and although many factors are needed to explain the variety and complexity of children's behaviors, certain consistencies in the results are apparent. Kaufman and Hollenbeck (1973) pointed out the similarity of many of the factors found in the preliminary factor analyses of the MSCA to those typically obtained from analyses of other test batteries such as the Stanford-Binet or WISC. This consistency may be further realized by relating the MSCA factors obtained in the present analysis at age 6+ to the primary mental abilities of school-age proposed by Thurstone and Thurstone (1941, 1953). A comparison of the six MSCA factors at age 6+ to the seven Thurstone factors that have been verified most frequently in research studies (Anastasi, 1968, pp. 329-330) reveals the following close correspondence between five of the abilities:

MSCA Factor at Age 6+		Corresponding Thurstone Factor	
I.	General Cognitive/Verbal	V.	Verbal Comprehension
III.	Perceptual-Performance	S.	Space
IV.	Memory	M.	Associative Memory
V.	Reasoning	I (or R).	Induction (or General Reasoning)
VI.	Quantitative	N.	Number

In addition, the sixth McCarthy factor at age 6+—II. Motor—is similar to the Motor (Mo) ability identified by the Thurstones and included in their battery to assess coordination of eye and hand movements (Thurstone and Thurstone, 1953).

Factor analyses at the preschool age levels, involving groups of substantial size, have been far less common than analyses of school-age children. Nevertheless, ample evidence has accumulated to show that at the ages of 2 or 3, and even at various levels of infancy, there are specific intellectual abilities that emerge as group factors—sometimes in addition to a general factor (McNemar, 1942; Quereshi, 1967; Richards and Nelson, 1939), and sometimes in the absence of *g* (Hurst, 1960; Meyers, Dingman, Orpet, Sitkei, and Watts, 1964; Stott and Ball, 1963).

An area of extreme interest to many investigators has been the relationship of the abilities at the preschool levels to those at school age levels and, in particular, the developmental progression of these abilities. A number of studies have explored this topic by conducting separate factor analyses of preschool and school age groups and comparing the resulting structures. However, these studies have not provided definite answers to the developmental question for a variety of reasons such as the following: the nature of the tasks was markedly different from age level to age level, even when the same instrument was used for all children (Stott and Ball, 1963); the age range of the different samples was not sufficiently broad, despite spanning preschool and school-age children (Hollenbeck and Kaufman, 1973; Osborne and Lindsey, 1967); the tasks assessed only a limited range of children's abilities due to the particular instrument studied (Quereshi, 1967) or to the specificity of the investigators' hypotheses (Meyers et al., 1964).

The present analyses are thus of great importance to developmental theory. First, the age span of 2½ to 8½ years is quite broad, covering a period of rapid and important behavioral change; moreover, the homogenizing influence of schooling is relatively minimal for children in this age range. Also, the nature of the test content is very similar throughout the age range. Although the wide age span necessitates

having a few tasks that are useful only for very young or older children within the range examined, most of the tests in the MSCA produced substantial variation in scores at virtually all of the standardization age levels (McCarthy, 1972, pp. 204-205). In addition, the tasks in the MSCA provide broad coverage of the cognitive abilities that have been shown to be important in many previous studies of children's intelligence, including Kelley's (1928) early factor analytic investigation. Finally, whereas many factor analytic studies in the literature utilized groups which were limited in size and were often from a particular city or socioeconomic class, the MSCA sample is representative of the U.S. population on many important variables such as color, region, and father's occupation.

### *Developmental Interpretation of the Results*

Certain developmental trends were quite evident in the present analyses. The most important was the emergence of one additional ability at each of three levels: age 3, age 5, and age 6+. First consider that the four abilities found at the youngest age level—Verbal, Motor, General Cognitive, and Memory—were also identified at every subsequent age level, with minor exceptions. (The exceptions: no General Cognitive factor at age 4; a blending of General Cognitive and Verbal factors at ages 5 and 6+.) Then, at age 3, a Perceptual-Performance factor appeared—and this factor too was isolated at each older age level. A Quantitative factor appeared for the first time at age 5 and recurred at age 6+, which agrees with Thurstone and Thurstone (1948) who found that the ability to work quickly and accurately with numbers emerges gradually from a more global quantitative ability that is not distinct from other factors (such as verbal) in the young child. Finally, a conceptual factor called Reasoning emerged at age 6+, probably reflecting the fact that many of the more difficult items in the MSCA are more abstract than the easier items intended primarily for preschool and kindergarten age children.

It is of interest to relate the present results to Garrett's (1946) hypothesis—i.e., that there is a large general intellectual ability in infancy and early childhood which diminishes in size as the child gets older due to a gradual differentiation of specific abilities. To consider Garrett's hypothesis, one must ignore the nonintellectual psychomotor tests, as Guilford (1967, p. 414) points out. In the present series of analyses, the gross motor variables tended to load only on the Motor factor, so this factor will not be considered in the following discussion.

The presence of the Verbal and Memory group factors as early as



age 2½—with the Verbal factor accounting for the same percentage of variance as the relatively small General Cognitive factor—provides evidence contrary to Garrett's hypothesis. Although the emergence of a new group factor at ages 3, 5, and 6+ seems supportive of Garrett's theory, it is apparent that these group factors did not result from successive differentiations of a large general ability. In addition, there were two main developmental trends working in the opposite direction from the Garrett hypothesis; namely, the blending of the Verbal group factor with *g* at the two highest age levels, which suggests that once a child reaches school age his general mental ability may be closely intertwined with his verbal facility; and second, the appearance of a Drawing factor at ages 3 and 4 but not in subsequent analyses. The net result of the important developmental trends is that there were approximately the same number of group factors at ages 3 through 6+, and that the percentage of variance accounted for by the several General Cognitive factors did not decrease with age. Overall, then, Garrett's hypothesis was not supported.

The clear Drawing factor that emerged at ages 3 and 4 is worthy of comment, because it may reflect the changing nature of the abilities required for successful performance on the drawing tests at different ages (rather than indicating an ability that "disappears" at age 5). One may hypothesize that design copying and drawing a child are predominantly motor tasks for younger children, and gradually become more conceptual once pencil-and-paper coordination is mastered. If the Drawing factors at ages 3 and 4 are perceived as involving fine motor coordination (rather than a specific cognitive ability), the results of the present analyses suggest that the drawing tests are primarily motor at age 3, partially motor and partially cognitive at age 4, and predominantly cognitive at ages 5 and above.

### *Construct Validity of the MSCA*

There are three main questions to ask regarding the factorial evidence of the MSCA's construct validity: (1) Are there factors which correspond to the abilities assessed by each of the six scales at some or all of the age levels studied? (2) Does the factor structure across age levels suggest that additional or alternative scales might have been selected for the MSCA? and (3) Do the factor loadings support the placement of each component test on its particular scale(s)?

The first question has already been answered in the affirmative, as a factor corresponding to each scale was identified in two or more of the analyses. A Motor factor was found in all five analyses, as were both Verbal and Memory factors. Of these three, Motor showed the least

age to age variation, which was also true in Richards and Nelson's (1939) analyses at ages 6, 12, and 18 months. The Verbal factor, however, was merged with General Cognitive at ages 5 and 6+, and the Memory factor accounted for less than 10% of the variance at these same two age levels. General Cognitive and Perceptual-Performance factors were extracted in four of the five analyses, and only the Quantitative Scale was represented in as few as two analyses. ■

In response to question #2, there does not seem to be evidence supporting the need for additional or alternative scales. Factors labeled Semantic Memory and Reasoning each appeared at only one age level. As mentioned earlier, the emergence of the Drawing factor at ages 3 and 4 suggests that Draw-A-Design and Draw-A-Child have a motor component for children in the younger half of the MSCA age range. (The meaningful loadings of Draw-A-Design on the Motor factors at ages 2½ and 3 also support this interpretation.) McCarthy's placement of the two drawing tests on both the Perceptual-Performance and Motor Scales reflects the dual nature of the abilities required for these tests. This scale structure seems as good as any alternative solutions that might be suggested by the present data—particularly when one considers the practical advantages of having the same set of scales for all children within the MSCA age range.

To answer question #3 regarding the appropriateness of the scale placement of the component tests, the pattern of factor loadings for each test was examined across the age range. The following points become apparent:

1. The Motor factor tended to be defined by the Leg Coordination, Arm Coordination, and Imitative Action tests, all of which are included on the *Motor Scale*.
2. Of the four tests on the *Memory Scale*, Pictorial Memory, Tapping Sequence, Part I of Verbal Memory, and Part I of Numerical Memory each had meaningful loadings on at least three of the five Memory factors. Of tests *not* on the Memory Scale, only Puzzle Solving and Word Knowledge—both of which require some recall of past experience—had as many as three meaningful loadings.
3. Of the three tests on the *Quantitative Scale*, Number Questions, Counting and Sorting, and Part II of Numerical Memory each had meaningful loadings on *both* of the Quantitative factors; no other variable had a loading of .25 or better on both of these factors.
4. Of the six tests on the *Perceptual-Performance Scale* (excluding Right-Left Orientation which is not administered to children

below 5) all except for Tapping Sequence had meaningful loadings on at least three of the four Perceptual-Performance factors. Of the tests *not* on the Perceptual-Performance Scale, only Counting & Sorting (which is largely nonverbal) and Verbal Fluency had three meaningful loadings.

5. Of the five tests on the *Verbal Scale*, Word Knowledge, Part I of Verbal Memory, Verbal Fluency, and Opposite Analogies had meaningful loadings on either four or all five of the Verbal factors. Among other tests, only Number Questions, Part I of Numerical Memory, and Conceptual Grouping had four meaningful loadings, and these bear a logical relationship to verbal ability.
6. Of the 15 tests which make up the *General Cognitive Scale*, 9 had meaningful loadings on three or four of the General Cognitive factors, and two others had meaningful loadings on two of the factors. Of the three gross motor tests not on the General Cognitive Scale, only Leg Coordination had as many as two loadings of .25 or greater.

Overall, then, the factor structure gave support to McCarthy's placement of the tests in the various scales, although in some instances the data might be interpreted as supporting alternative test placements. For example, Number Questions (which involves comprehension of oral questioning as well as number facility) might seem to merit placement on the Verbal as well as on the Quantitative Scale based on its loadings. As a second example, there is no empirical support for the placement of Part II of Numerical Memory on the Memory Scale. However, these and other similar discrepancies seem to be minor in the face of the high degree of consistency between the factor loadings and the scale structure at the various age levels. In addition, there are certainly logical and practical considerations which would support McCarthy's test placement (e.g., Part II of Numerical Memory is *logically* a memory test, and it would not have been *practical* to fragment the Numerical Memory test by placing only Part I on the Memory Scale). These inconsistencies are due to the inherent complexities of the mental processes of even simple tasks. The theoretical purity of factor structure that one strives for simply may not be attainable with a practical tool such as the MSCA.

The present analyses are basically consistent with the preliminary analyses of the standardization edition which were available when the scales were selected; nevertheless, there are also some important differences. For example, in the preliminary analyses at ages 3-3½, 5-5½, and 7½-8½, Leg Coordination did not have substantial loadings on the Motor factors; in addition, although Quantitative and

Perceptual-Performance factors were each tentatively identified in the previous analyses, they were isolated at *different* age levels (rather than at a single level, as was found in the present analyses at ages 5 and 6+). Thus, the factor analyses described herein not only give evidence for the construct validity of the MSCA; they are even more consistent with the final scale structure chosen by McCarthy than are the preliminary factor analyses.

## REFERENCES

- Anastasi, A. *Psychological testing* (3rd ed.). New York: MacMillan, 1968.
- Garrett, H. E. A developmental theory of intelligence. *American Psychologist*, 1946, 1, 372-378.
- Guilford, J. P. *The nature of human intelligence*. New York: McGraw-Hill, 1967.
- Hollenbeck, G. P. and Kaufman, A. S. Factor analysis of the Wechsler Preschool and Primary Scale of Intelligence (WPPSI). *Journal of Clinical Psychology*, 1973, 29, 41-45.
- Hurst, J. G. A factor analysis of the Merrill-Palmer with reference to theory and test construction. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1960, 20, 519-532.
- Kaufman, A. S. and Hollenbeck, G. P. Factor analysis of the standardization edition of the McCarthy Scales. *Journal of Clinical Psychology*, 1973, 29, 358-362.
- Kelley, T. L. *Crossroads in the mind of man: a study of differentiable mental abilities*. Stanford, Calif.: Stanford University Press, 1928.
- McCarthy, D. *Manual for the McCarthy Scales of Children's Abilities*. New York: The Psychological Corporation, 1972.
- NcNemar, Q. *The revision of the Stanford-Binet scale*. New York: Houghton Mifflin, 1942.
- Meyers, C. E., Dingman, H. F., Orpet, R. E., Sitkei, E. G., and Watts, C. A. Four ability-factor hypotheses at three preliterate levels in normal and retarded children. *Monographs of the Society for Research in Child Development*, 1964, 29, No. 5.
- Osborne, R. T. and Lindsey, J. M. A longitudinal investigation of change in the factorial composition of intelligence with age in young school children. *Journal of Genetic Psychology*, 1967, 110, 49-58.
- Quereschi, M. Y. Patterns of psycholinguistic development during early and middle childhood. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1967, 27, 353-365.
- Richards, T. W. and Nelson, V. L. Abilities of infants during the first eighteen months of life. *Journal of Genetic Psychology*, 1939, 55, 299-318.
- Stott, L. H. and Ball, R. S. *Evaluation of infant and preschool mental tests*. Detroit: Merrill-Palmer Institute, 1963.
- Thurstone, L. L. and Thurstone, T. G. Factorial studies of intelligence. *Psychometric Monographs*, 1941, No. 2.

- Thurstone, L. L. and Thurstone, T. G. *Examiner Manual for the SRA Primary Mental Abilities: Primary*. Chicago: Science Research Associates, 1948.
- Thurstone, L. L. and Thurstone, T. G. *Examiner Manual for the SRA Primary Mental Abilities for ages 5 to 7* (3rd ed.). Chicago: Science Research Associates, 1953.



## POSSIBLE SAMPLING BIAS IN GENETIC STUDIES OF GENIUS

DANIEL P. KEATING<sup>1</sup>  
University of Minnesota

The data from Terman's *Genetic Studies of Genius* (1925-1959) relating to sample size, mean IQ, and variance of IQ scores were analyzed in terms of their conformation to the theoretically projected statistics derived from a consideration of the normal curve. Deviations from the theoretical projections lead to the probable conclusion that the sample size was too small, with the IQ scores clustered more closely about a significantly higher mean than projected. Although the major findings of the "Genius" study are not cast into doubt by this analysis, caution is urged with respect to comparisons to a normal sample when the differences are not large.

THE five-volume *Genetic Studies of Genius* (1925-1959), edited by the late Lewis Terman, has been widely and justifiably acclaimed as a landmark in longitudinal research. Its refutation of myths widely held at the time (e.g., that highly intelligent children are weak and sickly, that early ability is rarely maintained through adolescence and into maturity) was a starting point for work with gifted children, as well as for much research into the intellectual development of individuals. The study also illustrated the many difficult and often intractable problems of large-scale longitudinal research, one of which is examined here more closely.

In Terman's (1925) selection of his gifted group, he realized that his sample was not entirely correct. He stated:

---

<sup>1</sup> The author would like to thank Julian C. Stanley for his assistance in the preparation of this note and also Melita H. Oden for her cooperation in supplying data from the Terman Study of the Gifted. This note was completed in conjunction with a project sponsored by the Spencer Foundation.

One may conclude that the method of selection employed, although far from ideal, probably led to the discovery of at least 80 percent and possibly 90 percent of all the cases who could have qualified in the school population canvassed (p. 33).

What he may not have realized was that his estimate of error might itself have been erroneous. There are a number of indications that it was.

This is most clearly seen from an examination of the normal curve. With a mean of 100 and a  $\sigma = 16$  (Terman and Merrill, 1937) the percentage of cases falling beyond  $+2.5\sigma$  (i.e., 140 IQ)  $\cong .62\%$ . Multiplying the population canvassed, 168,000 (Terman, 1925, p. 29), by this figure yields a projected sample of about 1042 cases over 140 IQ. Terman's (1925) actual yield was 649 cases (.38% of the population), or 61.22% of the projected sample. Further, the projected mean of the portion of the unit normal distribution beyond  $+2.5\sigma$  (140 IQ) is, after Kelley (1947, p. 297),

$$\begin{aligned}\mu_{>2.5\sigma} &= \frac{y_{2.5\sigma}}{1 - P_{<2.5\sigma}} = \frac{0.0175283}{0.0062097} \\ &= 2.8227\sigma,\end{aligned}$$

where  $y_{2.5\sigma}$  is the height of the ordinate  $2.5\sigma$  above the mean of the unit normal distribution, and  $P_{<2.5\sigma}$  is the area of the distribution below  $2.5\sigma$ .

The mean of this tail portion of the unit normal distribution is 2.82; thus the mean of scores beyond 140 IQ is

$$100 + (2.8227)(16) \cong 145,$$

where 100 is the overall IQ mean and 16 is the standard deviation of IQ scores. The actual mean of the gifted group was 151 (Terman, 1925, p. 45), a difference of 6; this is about  $.4\sigma$  above the mean of the normal distribution beyond  $+2.5\sigma$ .

Jensen (1969) has pointed out the variations in the normal curve for IQ at the extremes. This casts some doubt on the reliability of the difference between the projected mean and the actual mean. However, it reinforces the difference between the projected and actual size of the sample. The proposed alteration of the normal curve would, if anything, increase the percentage of area under the curve beyond  $+2.5\sigma$ , thus increasing the size of the projected sample.

The theoretical standard deviation for the normal distribution of scores beyond  $+2.5\sigma$  may also be calculated. The variance is, again after Kelley (1947, p. 298),

$$\begin{aligned}
 \sigma_{(>2.5\sigma)}^2 &= 1 + \frac{(2.5)y_{2.5\sigma}}{1 - P_{<2.5\sigma}} - \mu_{>2.5\sigma} \\
 &= 1 + \frac{2.5(0.0175283)}{0.0062097} - (2.8227)^2 \\
 &= 0.089187,
 \end{aligned}$$

where  $\sigma_{(>2.5\sigma)}^2$  is the variance of the unit normal distribution beyond +2.5 $\sigma$  (values derived from Pearson and Hartley, 1956). For  $\sigma = 16$ ,  $\sigma^2 = 256$ , and  $\sigma_{(>2.5\sigma)}^2 = 0.089187$ , the standard deviation for the portion beyond 140 IQ is

$$\sqrt{(.089187)(256)} = \sqrt{22.83187} \cong 4.8.$$

By comparison, the standard deviation of the obtained sample was 10.2 (Terman, 1925, p. 45). This suggests the picture of the theoretically projected sample having scores clustered much more closely around a significantly lower mean ( $p < .001$ ).

The schematic comparison of the distribution of obtained IQ's above 140 in Terman's (1925) sample with the tail portion of the unit normal distribution in Figure 1 demonstrates the nature of the discrepancies calculated above. The shaded area represents the discrepancy leading to the inflated mean of the actual sample. This indicates, as can be seen from Figure 1, that too few "low" subjects and too many "high" subjects were included in the sample.

The calculations of the sample were performed on the grouped data found in Terman (1925). There is the possibility that the calculations might be affected by the grouping procedure. Unfortunately the original ungrouped data is not directly recoverable (Oden, personal communication), but the appropriate calculations were performed on the ungrouped data which were easily recoverable. No important differences were found between the grouped and ungrouped statistics.

It might be argued that the above statistical arguments fail because of the nature of the 1916 revision of the Binet-Simon (Terman, 1916). The mean IQ and  $\sigma$  were not calculated at that time, and there was no specification of the IQ distribution as a normal curve. One might infer from the technical monograph which accompanied the 1916 Stanford revision (Terman et al., 1917, p. 43) that  $\sigma = 13.5$ . Using the same reasoning as above, we obtain the following statistics for the theoretically projected sample:  $N = 267$ ; mean = 143; standard deviation = 3.6. Thus the mean and standard deviation are even more discrepant from the obtained figures than with an assumption of  $\sigma = 16$ , but the obtained sample size is greater than the projected size rather than smaller.

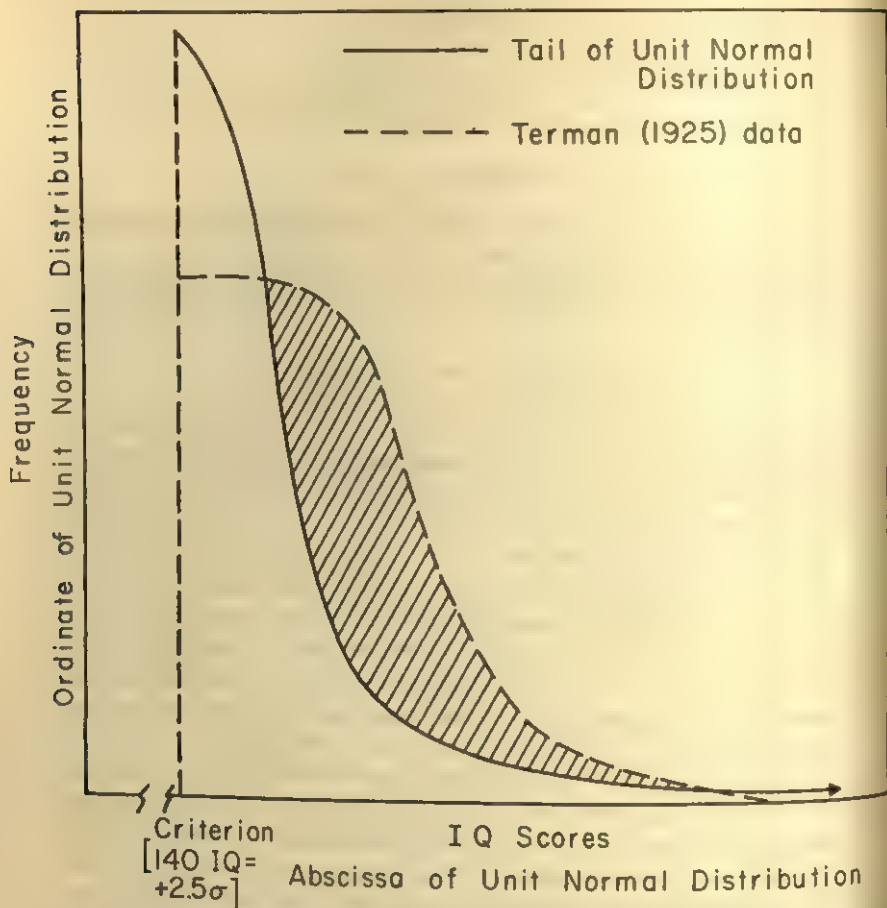


Figure 1. Schematic comparison of actual sample and theoretically projected sample.

There is also considerable deviation of the gifted group sample (Terman, 1925) from the sample projected by using Terman's (1916, p. 66) percentages. At least 0.55% of the standardization group score above 135 on the 1916 scale. Even allowing .05% for the 136-139 range, the projected sample size is 840. The actual sample (639) is thus 76.7% of the projected sample. The projected error, therefore, ranges from a low of nearly 24% to a high of almost 40%.

There are a number of plausible speculations regarding the source of this sampling error. First, the population from which the sample was drawn might have been markedly non-normal. Given the high number of students canvassed and the demonstrated normality of the

Stanford-Binet scores (Terman and Merrill, 1937), however, this seems unlikely.

The second speculation seems more likely. Taking together the facts of a too small sample, a (possibly) too high mean, and a (possibly) too large standard deviation, the intuitive inference is that too few of the "borderline" cases (140-150) were included. Terman (1925), after considering the defects of his selection techniques, conjectured about the nature of the cases he failed to locate:

They would almost certainly have been found a little less accelerated in school. Some would be excessively shy, others lazy, and still others lacking in adaptability [p. 33].

A third possible source of sampling error was the concentration on urban and suburban canvassing, with the rural population being nearly ignored (Terman, 1925, p. 29). The concentration of talent (as measured by IQ) tends to be greater in metropolitan as opposed to rural areas (e.g., Terman and Merrill, 1937; von Fieandt, 1958). The overload of "high" cases may be partly attributed to this factor.

If the error is actually closer to the 25-40% we have suggested than to the 10-20% Terman (1925) estimated, and if his characterization of those not included is correct, then one may easily see the ramifications for the significance of a number of this conclusions. Many of the statistically significant differences between his gifted group and the "general population" which were reported in 1925 and subsequently throughout the longitudinal study (Burks, Jensen, and Terman, 1930; Terman and Oden, 1947; Terman and Oden, 1959; and Oden, 1968) may in fact lack significance for the *specific* sample which Terman originally prescribed, especially since the "missed" cases were likely to be less differentiated from the general population than the cases in the obtained sample. It is clear from both the number and the degree of differences obtained that the major conclusions of the "Genius Study" were warranted. Caution is urged, however, in the interpretation of data from the "Genius Study" where only a small difference was reported.

## REFERENCES

- Burks, B. S., Jensen, D. W., and Terman, L. M. *Genetic studies of genius*. Vol. III: *The promise of youth*. Stanford, Calif.: Stanford University Press, 1930.
- Jensen, A. R. How much can we boost IQ and scholastic achievement? *Harvard Educational Review*, 1969, 39, 1-123.
- Kelley, T. S. *Fundamentals of statistics*. Cambridge: Harvard University Press, 1947.



- Oden, M. H. The fulfillment of promise: 40-year follow-up of the Terman gifted group. *Genetic Psychology Monographs*, 1968, 77, 3-93.
- Pearson, E. S. and Hartley, H. O. (Eds.). *Biometrika tables for statisticians*, Vol. 1. Cambridge, England: Cambridge University Press, 1956.
- Terman, L. M. *The measurement of intelligence*. Boston: Houghton-Mifflin, 1916.
- Terman, L. M., Lyman, G., Ordahl, G., Ordahl, L. E., Galbreath, N., and Talbert, W. *The Stanford revision and extension of the Binet-Simon scale for measuring intelligence*. Baltimore: Warwick & York, 1917.
- Terman, L. M. *Genetic studies of genius*. Vol. I: *Mental and physical traits of a thousand gifted children*. Stanford, Calif.: Stanford University Press, 1925.
- Terman, L. M. and Merrill, M. A. *Measuring intelligence*. Boston: Houghton-Mifflin, 1937.
- Terman, L. M. and Oden, M. H. *Genetic studies of genius*. Vol. IV: *The gifted child grows up*. Stanford, Calif.: Stanford University Press, 1947.
- Terman, L. M. and Oden, M. H. *Genetic studies of genius*. Vol. V: *The gifted group at mid-life*. Stanford, Calif.: Stanford University Press, 1959.
- von Fieandt, K. Psychological effects of urban and rural domiciles. *Acta Psychologica*, 1958, 14, 81-91.

## UNIVOCAL VARIMAX: AN ORTHOGONAL FACTOR ROTATION PROGRAM FOR OPTIMAL SIMPLE STRUCTURE<sup>1</sup>

DOUGLAS N. JACKSON AND HARVEY A. SKINNER

The University of Western Ontario

Univocal varimax is an orthogonal factor rotation strategy aimed at improving upon the simple structure qualities of a preliminary varimax solution. This is accomplished by targetting for patterned rotation the highest element in each row of the varimax factor loading matrix. This tends to yield a solution in which each variable in the final rotated matrix maximally loads on only one factor. Univocal varimax is particularly relevant to research problems in which each rotated factor should be marked by a relatively tight cluster of variables. A FORTRAN IV program is described for the efficient analysis of large input factor loading matrices.

THE purpose of this paper is to describe the essential characteristics of an orthogonal factor rotation program UNIVMX for achieving optimal simple structure.

The rationale underlying this procedure, termed univocal varimax, is an attempt to improve upon the simple structure qualities of a preliminary varimax solution so that each variable maximally loads on *only one* factor. That is, program UNIVMX seeks to orient each factor through a relatively tight cluster of variables. Factors in the final solution should be characterized by several high loadings, with the remaining loadings near zero. It is thus designed to avoid the situation, sometimes encountered in using standard varimax, of a single variable in which moderate loadings on each of several factors are obtained. In many respects, program UNIVMX may be considered an orthogonal

---

<sup>1</sup> The research for this paper was supported by the Defence Research Board of Canada Grant Number 9435-75 (UG) and Canada Council Grant Number S74-0761.

analogue to the Promax oblique rotation strategy of Hendrickson and White (1964).

Potential research applications include the Tucker and Messick (1963) points-of-view model of multidimensional scaling, where it is important that each point-of-view factor be marked by a definable subgroup of judges in the scaling experiment (Cliff, 1968). Another example is the use of *Q*-technique factor analysis for classification research (Skinner, Jackson, and Hoffmann, 1974). Each entity factor should be aligned with a relatively homogeneous cluster of individuals to facilitate interpretation of the solution. Indeed, univocal varimax may be used in any application in which a given test is expected to exhibit loadings on one and only one factor, e.g., multitrait-multimethod matrices involving correlations among measures of intellectual abilities or of personality constructs.

### *Computational Strategy*

Program UNIVMX first rotates an input factor loading matrix to a varimax criterion (Kaiser, 1958). Then, the varimax solution is scanned to identify on which factor each variable has a highest loading. An hypothesis matrix composed of 1, -1, or 0's is generated on the basis of two criteria: that a variable have a loading on a particular factor above a user-defined minimum (or a default value of  $|.50|$ ); and (b) that this loading is the highest loading for that variable. Finally, an orthogonal procrustean transformation is performed whereby the input factor loading matrix is rotated to a least-squares fit to the hypothesis matrix (Schönemann, 1966).

### *Input*

The data deck for each problem consists of (a) a title card, (b) a parameter card specifying the number of variables and factors comprising the input factor loading matrix, (c) a format card for reading the data, and (d) the input factor loading matrix.

### *Output*

Program output includes the (a) input factor loading matrix (optional), (b) preliminary varimax solution, (c) hypothesis matrix, (d) transformation matrix, and (e) final univocal varimax solution.

### *Capabilities and Availability*

The program is written in FORTRAN IV either for the CDC CYBER 73 or for IBM systems. The maximum input factor loading matrix is of order 200 variables by 30 factors. This limitation may be

easily modified by the user to handle larger or smaller matrices dependent upon system capacity. Instructions for use, and a source listing, which includes a sample problem, are available from Douglas N. Jackson, Department of Psychology, The University of Western Ontario, London, Ontario N6A 5C2, Canada. Requests for the program should specify which version, the CDC CYBER 73 or the IBM, is desired.

## REFERENCES

- Cliff, N. The "idealized individual" interpretation of individual differences in multidimensional scaling. *Psychometrika*, 1968, 33, 225-232.
- Hendrickson, A. E., & White, P. O. Promax: A quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology*, 1964, 17, 65-70.
- Kaiser, H. F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 1958, 23, 187-200.
- Schönemann, P. H. The generalized solution of the orthogonal procrustes problem. *Psychometrika*, 1966, 31, 1-16.
- Skinner, H. A., Jackson, D. N., and Hoffmann, H. Alcoholic personality types: Identification and correlates. *Journal of Abnormal Psychology*, 1974, 83, 658-666.
- Tucker, L. R., & Messick, S. An individual differences model for multidimensional scaling. *Psychometrika*, 1963, 28, 333-367.





## THE PATH ANALYSIS OF COMPLEX RECURSIVE SYSTEMS

CHARLES F. TURNER<sup>1</sup>

London School of Economics and Political Science  
University of London

Methods for the path analysis of complex (i.e., not fully recursive) causal models are briefly discussed. A computer program which simplifies analysis of such models and provides an option for automatically deleting marginal paths is described.

PATH analysis was introduced by Wright (1921; 1934) to provide geneticists with a methodology for systematically describing the functional relations between variables based upon their intercorrelations *and* theoretically derived assumptions of causal asymmetry. In Wright's own words, path analysis "is an attempt to present a method of measuring the direct influence along each separate path in such a [linear, additive, causal] system, and thus of finding the degree to which variation of a given effect is determined by each particular cause [1921, p. 557]." Recently it has been demonstrated that path analysis has a wide range of interesting applications in the causal modelling of social and psychological processes (e.g., Duncan, 1966; Hauser, 1972). Land (1969) has provided the interested reader with a recent and readable formulation of the basic theory and limitations of the method.

A computer program (PATHL) has been designed to simplify the path analysis of complex recursive systems in which variables that occur contemporaneously in the real world must be introduced simultaneously into a model. In the analysis of such models PATHL requires the assumption of noncausality within blocks of contemporaneous variables, as well as unidirectionality of effect between

---

<sup>1</sup> The Social Psychology Research Unit provided support for this work under a grant from the Social Science Research Council.

The author is now at the Department of Psychology, Columbia University.

Copyright © 1975 by Frederic Kuder

stages of a given model, and the other normal assumptions of path analysis concerning adequacy of specification, correlations among residuals, and homoscedasticity. Of course, the first assumption need not be made when PATHL is used in the analysis of a simple recursive system.

### *Method and Output*

PATHL initially computes the means, standard deviations, and a triangular matrix of product moment correlations for all variables in the database specified by the user. Subsets of this information are then selected at each stage of the modelling process. The resulting sub-matrices of intercorrelations are inverted by the Gauss-Jordon method to provide least squares estimates of the regression coefficients as well as standardized path coefficients. The path coefficient for the residual term is derived, and confidence statistics for the coefficients (*t*-values and standard errors) as well as an analysis of variance for the modelling stage are computed.

Following Duncan's suggestion (1966, p. 7), PATHL provides users with the option of automatically producing a *reduced* model in which paths of marginal absolute value or statistical significance are eliminated.

### *Machine Requirements*

PATHL is composed of routines written in FORTRAN IV and (IBM) OS/360 Assembler language. The execution of PATHL requires 90K bytes of core storage and two reusable i/o mediums.

PATHL, which has been successfully used on a variety of IBM/360/370 machines, should be convertible to other large-scale computing systems.

### *Execution Control and Data Management*

Any number of new databases may be constructed and analysed by PATHL. The user provides one statement of database parameters and one format statement to describe each new set of input data. Facilities are provided for data screening.

Models are built from endogenous and exogenous variables which the user specifies in a simple manner at each stage of model construction. No more than 40 variables may be used in any single model. There are no restrictions on the number of models which may be constructed from a given database subset.

### Timing

When executing the PATHL load module on an IBM/360-91, each modelling stage requires an average of less than one-tenth of a second of cpu time. Database manipulations are only required when new subsets are requested by the user; these manipulations increase execution time to an extent proportional to the frequency of their occurrence, the size of the database, and the efficiency of the storage format of the input data.

### Availability

The following materials may be obtained from the author: (a) source statements and optimized object modules stored in a convenient format on 9-track magnetic tape, (b) printed listing of the source code and exemplary output, (c) a detailed guide to the use of the program, and (d) technical notes for programmers wishing to implement PATHL on non-IBM machines.

### REFERENCES

- Duncan, O. D. Path analysis: some sociological examples. *American Journal of Sociology*, 1966, 72, 1-16.
- Hauser, R. Disaggregating a social psychological model of educational attainment. *Social Science Research*, 1972, 1, 159-188.
- Land, K. Principles of path analysis. In E. Borgatta (ed.), *Sociological Methodology: 1969*. San Francisco: Jossey-Bass, 1969.
- Wright, S. Correlation and causation. *Journal of Agricultural Research*, 1921, 20, 557-585.
- Wright, S. The method of path coefficients. *Annals of Mathematical Statistics*, 1934, 5, 161-215.



## IRIS: A COMPUTER-INTERACTIVE APL PROGRAM FOR RECOVERING SIMPLE ORDERS<sup>1</sup>

THOMAS J. REYNOLDS AND NORMAN CLIFF  
University of Southern California

IRIS, a computer-interactive APL program for developing a simple order for a set of stimuli, is described. IRIS executes interactively by presenting pairs to be judged at a remote terminal. The subject responds by indicating his preference, or, by indicating that he has no preference. From a subject's judgments, IRIS determines which judgments are implied via transitivity and presents only pairs for which no implications are known. A substantial reduction in number of paired-comparisons required to recover a simple order results.

THE purpose of this paper is to describe the rationale underlying as well as the major features of a computer-interactive APL program IRIS, the objective of which is to obtain an individual's simple ordering of a set of stimuli in as few paired comparison preference judgments as possible. In a general context, IRIS attempts to attack the number-of-judgment problem associated with theoretically sound comparative judgment models, namely, the requirement of  $N(N - 1)/2$  pairwise presentations constituting all possible pairs. This problem is considerable, in that, for even moderately large stimulus sets, the number of judgments required increases parabolically, reaching 300 and 780 judgments for  $N = 25$  and 40, far exceeding any reasonable expectation in terms of both a subject's time and attention span. IRIS attempts to resolve this problem by determining which pairwise presentations would be redundant, by implication, and thereby to avoid a duplication of information already known.

---

<sup>1</sup> Preparation of this material was supported in part by Public Health Service Grant MH 16474.

Copyright © 1975 by Frederic Kuder



INTERORD, a FORTRAN counterpart to IRIS, (using an identical rationale) has been developed and is currently available (Kehoe and Cliff, 1975).

A simple order may be completely determined by obtaining an observed preference judgment for each adjacent pair of stimuli (in the order). This principle of connectiveness in determining an order is founded in the graph theoretical assumptions of directed graphs (Harary, Norman and Cartwright, 1965). Thus, the complete determination of a simple order can be accomplished once these appropriate  $N - 1$  judgments connecting the adjacent stimuli in the order are known. The remaining  $\frac{1}{2}(N - 1)(N - 2)$  judgments become redundant, as they are implied via transitivity.

It has been proposed that  $\log_2 N!$  is the minimum number of judgments required to obtain the  $N - 1$  chain, which yields the complete order (Harary, Norman and Cartwright, 1965; Knuth, 1973). Since IRIS requires, for both errorless and errorful data, a number very close to this proposed minimum (Cliff and Reynolds, 1974; Reynolds, 1975), the number of pairs that need be presented to a subject when a simple ordering is desired is greatly reduced. For example, for  $N = 25$ , the minimum number is 84, a substantial saving in comparison to the 300 (all possible pairs). More significant, however, is the reduction in the number of required judgments for larger stimulus sets. In the  $N = 40$  example, where 780 represents the number of all possible pairs, the minimum number is 160, which remains well within a reasonable number of judgments to require of a subject.

The process of presenting only those pairs which are not redundant requires an interactive computer system utilizing a remote terminal from which the subject responds to each pair-wise presentation. Thus, the IRIS algorithm may be summarized, simply, as a process whereby a search for implications is made following each response, and, once having been determined, these implied judgments are recorded as if the subject made them directly. Thereby redundancy is avoided.

The search for implications is based upon a Boolean matrix multiplication procedure originally proposed by O'Neil and O'Neil (1973) and implemented by Cliff (1975, in press). The worth of this procedure stems from the fact that all elements of the response need not be subjected to matrix multiplication; rather, only those elements involved in the last paired-comparison need be considered in calculating the new implications. The saving in the total number of required calculations brought about by the utilization of this matrix multiplication short-cut not only accounts for the absence of any lag-time between presentations, but also keeps the cost of recovering an order from becoming financially prohibitive.

### *Program Features*

#### *Input*

Program input is from a remote terminal. The input includes subject identification, the number of stimuli, and the stimuli names. Available input options include: recovering a user-specified number of randomly selected pairs from which a validation index is calculated on the basis of the final order and specification of the initial pairs, instead of the default random assignment. Also included in IRIS is the possibility of allowing a no-preference response to a given pair, which is analogous to a tie or subject's indifference.

#### *Output*

Output is only available at the terminal. The output includes the recovered order of stimuli with their assigned ranks, a measure of predictive validity of the final order, and other pertinent run statistics, e.g., the number of judgments. Also available to be called out by the user are any of the matrices or vectors used in calculations including a summary in matrix form of all preference judgments made and their presentation number.

#### *Limitations*

A maximum of about 50 stimuli, depending upon options utilized, is possible for use.

#### *Computer and Program Language*

Written in APL, IRIS is currently implemented at an IBM 370/158 facility.

A copy of the program and sample output may be obtained from Norman Cliff.

### REFERENCES

- Cliff, N. Complete orders from incomplete data: interactive ordering and tailored testing. *Psychological Bulletin*, In press, 1975.
- Kehoe, J. and Cliff, N. INTERORD: a computer-interactive FORTRAN IV program for developing simple orders. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1975, 35, 675-678.
- Cliff, N. and Reynolds, T. J. A computer-interactive ordering program

for pairwise preference judgments. Paper presented at meeting of the Psychometric Society, Stanford, California, 1974.

Harary, F., Norman, R. and Cartwright, D. *Structural models: An introduction to the theory of connected graphs*. New York: Wiley, 1965.

Knuth, D. E. *The art of computer programing. V.3. Sorting and searching*. Reading, Mass. Addison-Wesley, 1973.

O'Neil, P. E. and O'Neil, E. J. A fast expected time algorithm for Boolean matrix multiplication and transitive closure. *Information and control*, 1973, 22, 132-138.

Reynolds, T. J. An interactive preference ordering model in APL and its Monte Carlo evaluation. Unpublished thesis, University of Southern California, 1975

## INTERORD: A COMPUTER-INTERACTIVE FORTRAN IV PROGRAM FOR DEVELOPING SIMPLE ORDERS<sup>1</sup>

JERARD F. KEHOE AND NORMAN CLIFF  
University of Southern California

INTERORD, a computer-interactive FORTRAN IV program for developing simple orders on a set of objects, is described. INTERORD executes in the interactive mode by presenting pairs to be judged at a remote terminal. The respondent at the terminal judges which one of the objects dominates the other. INTERORD utilizes the observed dominance relations to determine which additional relations are implied by the transitivity principle. Since pairs not implied are presented for judgment, a substantial reduction occurs in the number of judgments required in the pair comparisons procedure.

IN many situations investigators require the determination of a simple order on a set of objects. Examples include such tasks as having a subject order choice alternatives according to his or her preferences, having a personnel worker order job elements in terms of importance, or even having an individual order the cells of a factorial design with respect to a dependent variable of interest. Theoretically sound measurement procedures such as comparative judgment methods frequently require a very large number of judgments, many of which may seem redundant with respect to previous judgments.

### *Purpose*

The purpose of this paper is to describe principal features of a computer-interactive FORTRAN IV program for developing simple

---

<sup>1</sup> Preparation of this material was supported in part by Public Health Service Grant MH 16474.

Copyright © 1975 by Frederic Kuder

orders (INTERORD). Reynolds and Cliff (1975) have described an APL version of the interactive ordering procedure. The objective of INTERORD is to determine a simple order for an individual respondent from a minimum number of paired-comparison dominance judgments.

INTERORD operates in an interactive mode by presenting pairs of objects and by collecting dominance judgments from a respondent who is utilizing a remote control terminal. The interactive feature is required in order that the program may determine after each judgment which of the judgments not yet observed are redundant with respect to the previously observed judgments. INTERORD presents to the respondent at the terminal only pairs which are not redundant.

### *Rationale*

The determination of redundant pairs is based on the principle of transitivity of dominance judgments, which is fundamental to the construction of a simple order. The object pair  $(x, z)$  is redundant when it is implied by the transitivity principle. As soon as  $x$  dominates any object  $y$  and  $y$  dominates  $z$ , then  $x$  is assumed by transitivity to dominate  $z$ . Although INTERORD does not present for judgment such implied pairs, it does assume rather that the implied judgment for that pair exists and does yield only those pairs for which no judgment is implied. The algorithm for determining these transitive implications is based on matrix multiplication utilizing Boolean arithmetic (Cliff, 1975).

A simple order is completely determined when an order of objects can be found in which each pair of consecutive objects is connected with an observed dominance relation. When such an order exists, all other pair relations must be implied by transitivity. It is not necessary to present them for judgment. INTERORD relies on this principle in determining the simple order for the respondent. This principle of connectedness is derived from a graph theoretic approach to relations among objects (Harary, Norman, and Cartwright, 1965). The number of judgments required to define a complete simple order has a theoretical minimum equal to  $\log_2 N!$  where  $N$  is the number of objects. It has been found in practice that this number (which is 62 for 20 and 215 for 50) is a very close approximation to the actual number of judgments required by INTERORD.

The program also gathers a number (user-specified) of extra judgments which are stored separately. After the interactive session, these judgments are used to provide validation data for the final order.



### *The Program*

#### *Input*

Program input is from the remote terminal. The input includes respondent identification, the number of objects, the object names, an advantageous starting order if appropriate, the number of randomly selected extra judgments, and some output options. An option exists for the object names to be read from a previously created data set instead of from the terminal.

#### *Output*

There are three output modes. Output available at the terminal includes run statistics, object names, and the starting and final order. Printed output consists of run statistics, a measure of the predictive validity of the final order (utilizing the random extra judgments), the judgments, object names, the starting and final orders, and an optional matrix of responses. Punched card output is similar to the terminal output. Printed and punched output are created by an accompanying but separate program which executes in batch mode. INTERORD writes the run data onto a user-created data set which the output program in turn reads.

#### *Limitations*

Maxima of 50 objects, 240 judgments, and 30 random extra judgments are allowed.

#### *Computer and Program Language*

INTERORD and the output program are written in FORTRAN IV H compiler for the IBM 370/158. It should be noted that 66K is required for INTERORD and that 50K is needed for the output program.

A copy of the program and sample output may be obtained from Norman Cliff, Department of Psychology, University of Southern California, Los Angeles, California 90007.

#### REFERENCES

- Cliff, N. Complete orders from incomplete data: Interactive ordering and tailored testing. *Psychological Bulletin*, In press.

Harary, F., Norman, R., and Cartwright, D. *Structural models: An introduction to the theory of connected graphs*. New York: Wiley, 1965.

Reynolds, T. J., and Cliff, N. IRIS: A computer-interactive APL program for recovering simple orders. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1975, 35, 671-674.

## MAPPING INDIVIDUAL LOGICAL PROCESSES

FREDERICK O. SMETANA

North Carolina Science and Technology Research Center  
Research Triangle Park, N. C.

A technique to measure and describe concisely a certain class of individual mental reasoning processes has been developed. The measurement is achieved by recording the complete dialog between a large, varied computerized information system with a broad range of logical operations and options and a human information seeker. A type of flow chart, familiar to computer programmers, is used to delineate this dialog in a hierarchical fashion. An example of such a chart is given in the paper. Results obtained on a limited number of investigations suggest the technique may be useful for studying various aspects of mental performance.

CERTAIN individuals display an unusual ability to find specific information quickly in large collections such as libraries. Perhaps if one understood sufficiently well the complex logical process by which these individuals operate, one might quantify the processes somewhat in the form of a logical "map." Then, by obtaining and analyzing a statistically significant number of such logical mappings, one may be able to deduce certain universal laws for effective information transfer.

The prospect of being able to do this task and the knowledge of its potential impact on education led, upon the acquisition of suitable measuring equipment, to the design of a brief experiment whose results are reported in this paper. The experiment is described in detail in Smetana (1975).

Briefly, the experiment recorded the complete dialog between a computerized information retrieval system and a human information seeker. Since the information collection is large and varied ( $7 \times 10^6$  citations in 34 broad subject categories) and since the range of logical operations and options which one can request that the system perform

is quite varied, the information seeker can permit his imagination and experience almost limitless freedom in retrieving answers to specific questions. In other words, the individual's thought processes are not significantly constrained by the system. By recording the instructions given by the information seeker to the machine, its response, and his reaction to that response, it is possible to gain considerable insight into what the seeker is actually thinking during the search process—literally how his mind works in solving problems of this type. This insight can be translated into a formal flow chart which depicts in symbolic shorthand the chronological and hierarchical relationships between concepts, responses, and instructions.

On the basis of some limited but successful experience with these recording and charting techniques it is suggested that they may be useful to other investigators attempting to measure and to characterize various aspects of mental performance.

### *Recording Device*

Instructions to the computer are entered by means of typewriter keyboard. The instructions and the system response are displayed on a TV screen placed behind the keyboard. To one side of the keyboard is a printer which records everything displayed on the TV screen. These typed records then are the raw material from which the flow charts are constructed.

### *Flow Charts*

Construction of the flow charts (an example of a section of a complex flow chart is shown in Figure 1) requires the exercise of some judgment on the part of the drafter, particularly concerning the antecedents or sources of certain concepts. Thus there may be some variability in assigning hierarchical relationships. But if one has a reasonable familiarity with the subject area, this variability is usually not significant; in fact, it will usually be found that flow charts drawn by various individuals with an understanding of the subject area and of the approach to be used in constructing the charts seldom differ.

As an example of this lack of any practical difference in basic perception, one may mention that the 10 flow charts assembled to date were drafted by three different individuals. Yet when the one individual examined the flow charts drafted by the other individuals and compared them with the dialog recording, he seldom found an outright error and, more importantly, seldom could offer a suggestion for improving the graphical representations. This circumstance indicates that the process is reasonably independent of the drafter.

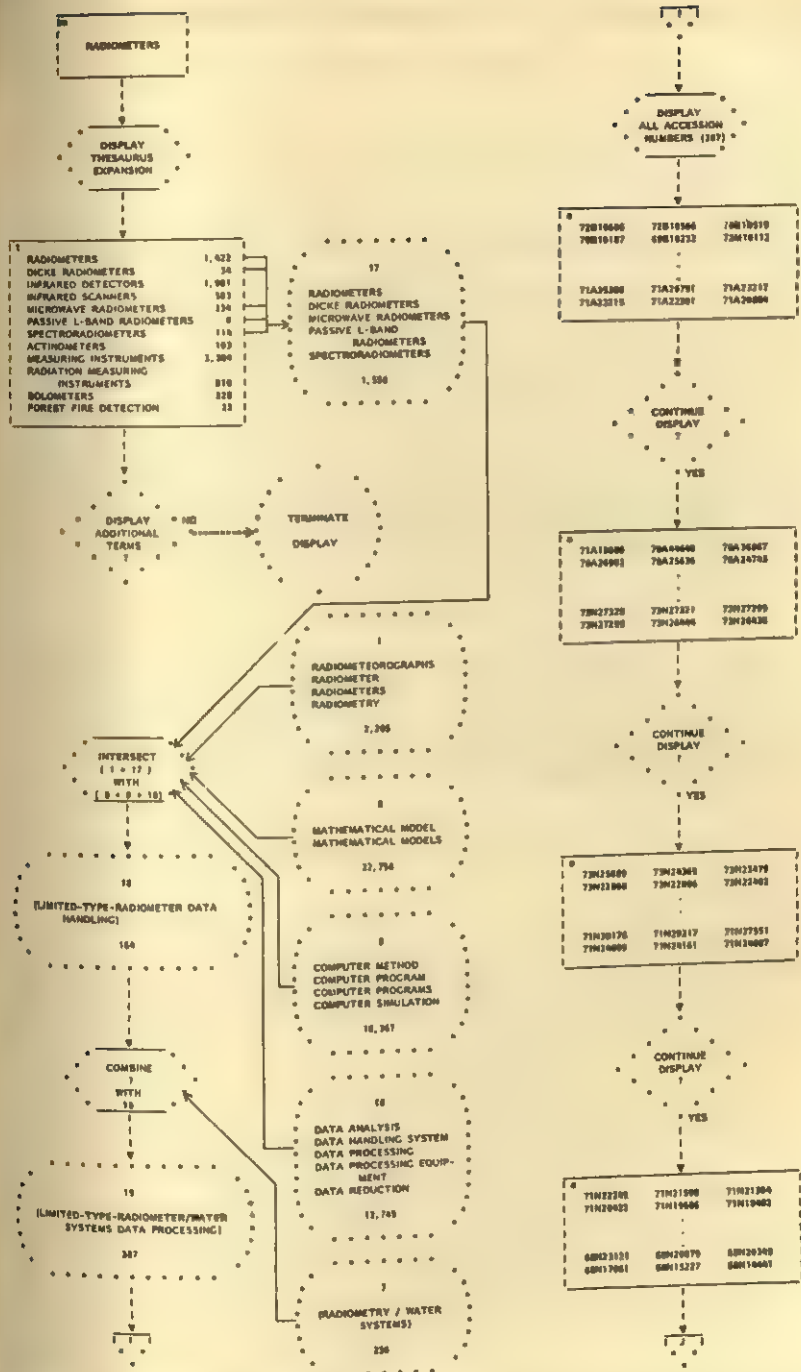


Figure 1. Example of a section of a flow chart depicting the hierarchical relationships in the dialog between a human information seeker and a computerized information system.



### *Concluding Remarks*

Examining a flow chart in detail is a fascinating experience in retracing someone's thought process. It becomes even more fascinating when one examines how several individuals have attacked the same problem. One soon discovers that, for complex problems at least, it is very difficult to determine (even after the fact) what was the optimal strategy for solution of the problem. That several completely different yet seemingly equally logical approaches have achieved comparable results indicates the need to analyze a large number of solutions for statistical significance in order to detect which approach or class of approaches has achieved superior results for a particular problem. The technique of recording the thought process dialog in detail and of expressing it in the form of a flow chart seems particularly well suited to characterizing the human thought process in a formal manner so that it may serve as an ideal raw material for further analysis.

### REFERENCE

- Smetana, F. O. Mapping individual logical processes in information searching. North Carolina Science and Technology Research Center Technical Report STR-506, November 1974. 22 pp. Available from the National Technical Information Center as a NASA Contractor's Report 142207, N75-19074.

## ITANA—III: A FORTRAN IV PROGRAM FOR MULTIPLE-CHOICE TESTS AND ITEM ANALYSIS

BARUKH NEVO, ELI SHOR, AND RACHEL RAMRAZ  
University of Haifa, Israel

A two-phase FORTRAN IV program called ITANA—III for an IBM 1130 computer is described that permits computation of psychometric characteristics of multiple-choice examinations including test statistics (phase I) and item statistics (phase II). Consisting of 280 statements, the program can handle up to 200 items with not more than 9 alternatives from samples of examinees not exceeding 2000.

THE purpose of this paper is to describe a two-phase computer program (ITANA—III) that was designed to calculate various psychometric characteristics of (a) multiple-choice tests (phase I giving test statistics) and (b) individual items (phase II providing item statistics).

The results of the analysis of scores on a multiple-choice test carried out in phase I may be used in applying and interpreting the results of item analyses carried out in phase II. Written in FORTRAN IV the program is for an IBM 1130 computer.

### *Program Input and Output*

The input consists of the answers of  $N$  subjects to a multiple-choice test with  $m$  items, each with  $k$  mutually exclusive alternatives.

The output of phase I consists of (a) individual raw scores, (b) mean and standard deviation, (c) frequency distribution (table, histogram), (d) chi-square test for testing the normality of the frequency distribution, (e) split-half reliability coefficient between scores on odd and

even items, (f) Kuder-Richardson reliability coefficient (formula 20), and (g) product-moment correlation with one or more external criteria (if supplied).

The output of phase II consists of (a) point biserial correlation over subjects between individual items and raw scores, (b) point biserial correlation over subjects between individual items and one or more external criteria (if supplied), (c) difficulty-levels of individual items ( $n/N$  or  $n/A$ , where  $n$  indicates the number of subjects who answer correctly this item,  $N$  indicates the total number of subjects, and  $A$  denotes the number of subjects attempting to answer this item), and (d) proportion of answers for each of the  $k$  alternatives.

### *Program Capacity and Limitations*

The program consists of 230 statements. Limitations are:  $N \leq 2000$ ;  $m \leq 200$ ;  $k \leq 9$ .

### *Availability*

A listing of FORTRAN statements, detailed directions for use, and example input and output may be obtained from Barukh Nevo, Department of Psychology, University of Haifa, Mt. Carmel, Haifa, Israel.

## A PROGRAM SYSTEM FOR THE ESTIMATION OF CHARACTERISTICS OF THE TEST SCORE DISTRIBUTION RESULTING FROM TEST ITEMS WITH GIVEN STATISTICS

LEE L. SCHROEDER  
Burlington County College

This paper describes a program system which was developed for the purpose of making estimates about the nature of the distribution of test scores which will result from an administration of a test composed of items on which estimates of item-difficulty and discrimination are available. In developing these estimates, the test constructor is free to make assumptions about the nature of the group of examinees and the number of examinees to be tested. It is expected that test developers will find this program system useful in the test development process.

AN integral step in the test development process is item pretesting. After test items are written, critiqued and revised, it is common to assemble preliminary tests composed of these items and to administer them to groups of subjects representative of the population for which the test is designed. Subsequently, the items are statistically analyzed to uncover any flaws in the items which were not formerly apparent. Items surviving this screening then become part of the pool from which items will be drawn to assemble the final test.

Many standardized tests are revised annually. The annual test versions are assembled based on item statistics determined from item pretesting as just described. Whenever tests are assembled from item pools in this manner, one problem is common for all test developers. That is, since the items on a new version of an examination were not necessarily tested *together*, it is necessary to develop methods of estimating the nature of the distributions of scores which will result from the administration of the test version. The only parameter of a score

distribution which can be estimated directly is the mean. The familiar formula

$$\bar{X} = \sum_{i=1}^n p_i$$

indicates how the test mean may be estimated from prior knowledge of the difficulty levels of the items. In this formula,  $\bar{X}$  is the test mean,  $n$  is the number of items on the test, and  $p_i$  is the difficulty level, or proportion correct, for the  $i$ th item. Several other parameters of the score distribution, however, are of interest. Among these parameters are the standard deviation of scores and the estimated test reliability (KR-20).

### *Overview of the Computer Programs*

To facilitate the review of expected sampling distributions of score statistics based on prior knowledge of the item-difficulty level and item-test biserial correlation for each item in an item set, the Test Simulation System (TSS) was developed. TSS was developed in the BASIC language to operate in a time-sharing mode. The system consists of five program segments and of several data files. Only one program is required to be in core at any moment so as to minimize core requirements.

Using TSS, the test constructor input may specify the number of subjects to be included in the simulation as well as the number of complete test simulations to be made. Since the simulation scheme is based on the latent trait model (Baker, 1965), the mean and standard deviation of the trait distribution are specified by the user. (A mean of zero and standard deviation of one are assumed as in score form, standard.) Last, the item difficulty levels ( $p$ ) and item test biserial correlations ( $r$ ) are entered as prompted by the computer.

Subsequently, the system output completes as many test replications as indicated for as many subjects as indicated. For each item,  $p$  and  $r$  values are converted into parameters of the adjusted cumulative logistic function (Lord and Novick, 1967) and, for each subject, a probability of scoring correctly on the item is generated. Based on these probabilities, all subjects are scored on each item. The output of this process is a test matrix. This test matrix is then analyzed by computing the twenty-two statistics in Table 1. These twenty-two statistics are stored, for each replication, in a disk file. After all replications have been performed, the final program segment produces histograms and summary statistics for each of the twenty-two statistics. From these histograms, or expected sampling distributions, for each statistic, inferences may be drawn regarding the nature of the score distribution which would result if the items perform as anticipated.



TABLE I  
*Twenty-Two Test and Item Statistics Associated with Test Matrix*

Sample	Mean	Statistics		
		Standard Deviation	Skewness	Kurtosis
Test scores	1 <sup>a</sup>	2	3	4
Item-Difficulty Levels	5	6	7	8
ETS Delta	9	10	11	12
Item-Test Point-Biserial Correlation	13	14	15	16
Item-Test Biserial Correlation	17	18	19	20
		21 Kuder-Richardson 20		
		22 Kuder-Richardson 20 adjusted by the Spearman-Brown formula to a 100-item test.		

<sup>a</sup> The number in the Table is the number of the variable in the system output.

### Discussion

When applied to a post-hoc analysis of a teacher made test, TSS was shown to provide exceptionally accurate estimates of parameters which were observed in the test data (Schroeder, 1975). Furthermore, Schroeder (1975) has shown the system to be valid in that known relationships between test item and test score statistics based on empirical research are found to exist in data generated from TSS.

### Availability

Program listings and a paper describing the use of TSS in research and test construction activities may be obtained from Dr. Lee L. Schroeder, Director of Measurement and Evaluation, Burlington County College, Pemberton, New Jersey 08068.

### REFERENCES

- Baker, F. B. Origins of the item parameters  $X_{00}$  and B as a modern item analysis technique. *Journal of Educational Measurement*, 1965, 2, 167-178.
- Lord, F. M. and Novik, M. R. *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1958.
- Schroeder, L. L. *An investigation of the relationships between selected test item and test score statistics based on simulated mental test data*. Ann Arbor, Michigan: Xerox University Microfilms, 1975.



## A COMPUTER PROGRAM TO CALCULATE ADJUSTED AND UNADJUSTED INTERRATER RELIABILITIES FOR SETS AND SUBSETS OF JUDGES

JOHN F. GREENE

University of Bridgeport

WILLIAM M. McCOOK

University of Connecticut

FRANCIS X. ARCHAMBAULT

Abt Associates

A computer program to be used to assess interrater reliabilities has been written. Given judges' ratings on a set of variables pertaining to subjects or events, the program will produce for each variable, a printed analysis of variance summary table for between subjects/events, within subjects/events, and between judges sources of variability, a reliability coefficient, an adjusted reliability coefficient and means and standard deviations for each rater. Further, analysis of variance and reliability output for subsets of judges is generated.

WITH many psychometric instruments, particularly those that elicit responses from open ended questions, the judgment of the scorer is a critical element. Examples of such instruments include tests of creativity and projective tests of personality. It has been noted that with these types of instruments there is as much a need for an estimate of scorer reliability as there is for more conventional reliability coefficients (Anastasi, 1968, p. 86).

In response to the need to assess between judge variability and subsequently interscorer reliability, a cycling type of computer program has been written. The program generates a reliability estimate of the pooled ratings of the judges using analysis of variance techniques (see

Winer, 1962, pp. 124-132) for all judges and for combinations of subsets of judges. At present, the program cycles through all combinations of one less than the total number of judges in addition to the total set of judges. The user may find this procedure useful if he needs to determine whether any one of the judge's ratings should be deleted because of systematic variation in that judge's scores that can be related to differences in that judge's training, experience, or frame of reference (see Greene, 1970, p. 26). The program also generates adjusted reliabilities, generally higher than the original reliabilities, which have empirically eliminated the effect of differences in judges' means. These reliability estimates should be utilized when the investigator is not willing to accept the assumption of mean homogeneity (Ebel, 1951). Winer (1962, p. 128) refers to this procedure as adjusting for differences in frame of reference.

### *Input*

Input for the program consists of the following parts:

1. A control card describing the number of variables that were rated, the number of judges, and the number of subjects.
2. A data matrix of each judge's ratings for each subject on each variable.

Data are read in for each subject, one judge at a time, for all of the variables under consideration.

### *Output*

The program yields the following outputs for each variable:

1. An analysis of variance Summary Table which includes sources of variation for between subjects/events, within subjects/events, between raters, residual and total, as well as corresponding degrees of freedom, sums of squares, and mean squares.
2. An estimate of the interrater reliability based upon the pooled ratings of the judges.
3. An adjusted reliability coefficient based upon elimination of the effect of differences in rater means.
4. A mean and standard deviation for each of the judge's ratings.
5. Separate analysis of variance, interjudge reliabilities, and adjusted reliabilities for each subset of judges. A judge code parameter indicates which judge was not considered in the particular analysis. Judge code 1 indicates that all judges were considered in the analysis.

6. A summary table of reliabilities and adjusted reliabilities for each variable and each judge code.

### *Capabilities and Limitations*

The program is written in PL/1. Central processing unit time was approximately 21 seconds and required 105K with 5 judges, 4 variables, and 137 subjects on an IBM 360, Model 65 computer running under OS. The program is presently set up to handle a maximum of 5 judges, 200 subjects, and 23 variables but is easily modified to handle larger data sets and different variations of subsets of judges.

### *Availability*

A listing of the program, sample problem, and documentation is available from Dr. William M. McCook, University of Connecticut, School of Pharmacy, Box U-92, Storrs, Connecticut 06268.

### REFERENCES

- Anastasi, A. *Psychological testing*. (3rd ed.). New York: The Macmillan Company, 1968.
- Ebel, R. L. Estimation of the reliability of ratings. *Psychometrika*, 1951, 16, 407-424.
- Greene, J. F. Scoring creativity tests by computer simulation. Unpublished doctoral dissertation, University of Connecticut, 1970.
- Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.





## A PROGRAM FOR THE T-SCORE NORMAL STANDARDIZING TRANSFORMATION

RONALD C. WIMBERLEY  
North Carolina State University

This FORTRAN program transforms raw values of a variable into T-scores having a normal distribution with a mean of 50 and a standard deviation of 10. Options are available for the categorization of data, the assignment of missing data, and the merging of old input data records with new output records which contain T-scores of the variables for each case. New records alone or new plus old record output may be placed on cards, tape, or disk.

THIS article describes a program for the T-score technique of normal standardization. T-scores transform a raw score distribution, regardless of its skewness or kurtosis, into a normal distribution with a mean of 50 and a standard deviation of 10. First, this technique places a set of input raw scores into a cumulative distribution. Then, the raw scores are assigned to the Z-scores which correspond to the Y-ordinates of either the means or the midpoints of the original raw score categories. These Z-scores are next transformed into T-scores ranging from 00 to 99 by the formula,  $T = 10Z + 50$ . One discussion of this technique is given by Walker and Lev (1958, pp. 192-201). This technique is not to be confused with other so-called T-normalizing approaches which calculate the underlying Z-scores directly from a raw score distribution without adjusting them to their proper proportions under the normal curve.

---

<sup>1</sup> This program was partially supported by the North Carolina State University Faculty Research and Professional Development Fund as well as the North Carolina State University Agriculture Experimental Station Regional Project NE-89. Appreciation is expressed to Edward Cureton for showing the writer the merits of the T-score transformation and to Alen Baker for programming assistance.

### *Rationale*

This T-score program is useful for approximating the normality assumptions of linear parametric statistics when ordinal data are used. An inherently linear relationship among the T-scores of different variables is free of mismatched kurtoses, skewnesses, and standard deviations which attenuate correlations or which lead to artificial nonlinearities in regressions. Furthermore, the T-score transformation should generally result in a more nearly normal distribution than that provided by other transformations such as those from logarithms, exponents, or roots.

### *Description and Input Capacities*

The program which is written in FORTRAN IV, H-level, is for use on an IBM 370/165 system. Input may be from cards, tape, or disk. The program comes in two versions. The first, REGLR NORMSTAN, is for as many as 120 variables which have as many as 120 values apiece. The other, SUPER NORMSTAN, is for 10 variables, any one of which may have up to 1000 values. Since calculation of T-scores from a large number of values is quite time consuming, the program allows variables to be categorized. Forty-one or fewer categories may be specified by the program user.

### *Computation Strategy*

Raw variables in either REGLR or SUPER NORMSTAN which have six or fewer categories are converted to T-scores by the mean technique. This subroutine places each raw score value at the Y-ordinate for the mean of a category's cumulative proportion under the normal curve. For variables with seven or more categories, the midpoint technique is used to place each raw score at the median rather than at the mean of its category in the cumulative proportions. The mean technique provides a better approximation to the normal curve than does the midpoint technique when there is a small number of raw score categories.

Missing data may be excluded from the calculation of T-scores by a subroutine for assigning missing data codes to another two-digit number. If a raw score variable has several missing data categories, such as "not applicable" and "no answer," these may be given new scores, such as 98 and 99 respectively, or both may be awarded the same score. These extreme reassignment values are unlikely to occur in an actual set of T-scores, since they are nearly five standard deviations

from the mean. If, instead, it is desired to reassign missing data to the mean of the T-score distribution, the designated value would be 50.

### *Output*

Output of the T-scores may be merely a printout or a printout plus either card, tape, or disk output of new data records containing case identification numbers, record numbers, and a series of two-digit T-scores for the transformed variables. Identification numbers are copied from input data records. The record number may be incremented from the number of input data records or specified by the user. The T-scores begin in column 13. Should there be more than 34 variables, their T-scores are continued to another record bearing the same case identification and an incremented record number. In addition, the new record(s) for each case can be produced by themselves or merged with reproductions of the input case records on cards, tape, or disk.

### *Availability*

For further information about this program or on obtaining copies of it, readers may communicate with the author at North Carolina State University, Post Office Box 5535, Raleigh, North Carolina 27607.

### REFERENCE

Walker, H. M. and Lev, J. *Elementary statistical methods*. New York: Holt, Rinehart and Winston.





## DISTINGUISHING BLANKS FROM ZEROS IN FORTRAN ON THE IBM 360 COMPUTER

NATHAN JASPEN  
New York University

A method is presented for distinguishing between blank fields and zeros in FORTRAN programs written for the 360 Computer. The method utilizes the T notation.

WHEN scores are missing from some subjects in a score data matrix, the researcher must be careful that the computer does not treat the missing data, represented by blanks in the card deck, as zeros. While blanks and zeros may be treated alike to obtain a sum, they must be distinguished from each other if means, standard deviations, or other summary statistics are computed. This paper presents a method for distinguishing blank fields from zeros.

Neither the integer nor the floating point notation in FORTRAN distinguishes between a blank field and a zero. The alphabetic format does distinguish, but of course alphabetic characters are not quantities, and arithmetic operations cannot be performed upon them. Thus, if the quantities 123 and 246 are read from a data card subject to the format expression (213), they can be added together, but if they are read subject to the format expression (2A3), they cannot be added together.

The T format code specification (T for tabulate) was developed primarily for the printing of column headings, but it may also be used on the input to skip from one column to another. The skipping can be backwards as well as forwards; hence the programmer can attain the same effect that a REREAD statement has on other computers. For instance, the format expression (80I1, T1, 80A1) reads all 80 columns in the data card twice, first in the integer mode and then in the alphabetic mode.

In the appended program, the critical statement is the format statement labelled 2. The T notation causes the scores, which have already been read as alphabetic fields, to be re-read as floating point quantities.

Five score cards, shown in Appendix 11, furnish the input data for this program. In this program a PRINT statement follows each READ statement; hence all input cards are printed. After the last data card is read, the three sets of computations (the sums, the  $N$ 's, and the averages) are printed. Each  $N$  includes zero scores, but excludes blanks.

### *Limitations*

A restriction of this method of expressing data in more than one mode is that the characters must be legal in each mode. For example, it is not legal to use the FORMAT expression (A3, T1, F3.1, T1, I3) to read the quantity 1.2, because the quantity 1.2 is not an integer and cannot be expressed in the I notation.

### Appendix 1

#### A FORTRAN Program that Distinguishes Zeros from Blanks

```

C   TO DISTINGUISH BLANKS FROM ZEROS
C   N JASPEN MARCH 1974
C
      DIMENSION A(5), X(5), S(5), SUM(5), AV(5)
      DATA BLANK
1     FORMAT (A1)
2     FORMAT (2X, I6, 5A4, T9, 5F4.0)
3     FORMAT (2X, 'SUM', 3X, 5F4.1)
4     FORMAT (2X, 'N ', 3X, 5F4.1)
5     FORMAT (2X, 'AV ', 3X, 5F4.1)
      DO 50 I = 1, 5
        S(I) = 0
        SUM(I) = 0
        AV(I) = 0
50    CONTINUE
100   READ (5, 2, END = 300) ID, A, X
      PRINT 2, ID, A
      DO 200 I = 1, 5
        IF (A(I)-BLANK) 150, 200, 150

```

```

150  S(I) = S(I) + 1.0
      SUM (I) = SUM (I) + X (I)
200  CONTINUE
      GO TO 100
300  DO 400 I = 1, 5
      IF (S(I)) 400, 400, 350
350  AV (I) = SUM (I)/S (I)
400  CONTINUE
      PRINT 1, BLANK
      PRINT 3, SUM
      PRINT 4, S
      PRINT 5, AV
      CALL EXIT
      END

```

## Appendix 11

## Printed Output of the Program in Appendix 1

11111	1	2	3	2	
11112		2	3	0	
11113	2	4	1		
11114	6		1		
11115		0	1		0
SUM	9.0	8.0	9.0	2.0	0.0
N	3.0	4.0	5.0	2.0	1.0
AV	3.0	2.0	1.8	1.0	0.0



## THE CALCULATION OF CORRELATION MATRICES USING SINGLE SUBSCRIPT NOTATION

NATHAN JASPEN  
New York University

A method is presented of calculating correlation matrices using single subscript rather than double subscript notation. This saves time and space, and permits the calculation of larger matrices in the space available.

CORRELATION programs usually use double subscript notation, since the correlation matrix is two-dimensional. The purpose of this paper is to present a method of producing correlation matrices that employs single subscripts. The single subscript method is superior to the double subscript method because it saves time and space, and permits the computation of large matrices.

### *The Double Subscript Method*

Coding such as the following is typical for calculating sums of cross-products:

DO 500 I = 1, M

DO 500 J = 1, M

500 C(I, J) = C(I, J) + X(I) \* X(J)

In this illustration, M is the number of variables, X represents the variables, and C represents the cross-products. It is assumed that M has already been defined for the particular problem and that the cross-product matrix C has been properly zeroed. The matrix C is square, consisting of M columns and M rows, and all cells are computed.

An alternate method of coding, probably more widely used, is the following:



$$\text{DO } 500 \text{ I} = 1, \text{ M}$$

$$\text{DO } 500 \text{ J} = 1, \text{ I}$$

$$500 \text{ C(I, J)} = \text{C(I, J)} + \text{X(I)} * \text{X(J)}$$

This triangle method requires only about half as many calculations as the square method, and therefore, uses approximately half as much time. The space requirements are, however, identical if double-subscript notation is used. Furthermore, the maximum size cross-product matrix that can be squeezed into any given number of storage units is an M by M matrix, where M is the square root of the number of storage units available for the crossproduct matrix.

### *The Single Subscript Method*

Either the square method or the triangle method can be executed with single-subscript notation, but only the triangle method is of interest here. The advantage of the single subscript notation is that it saves not only time but also space. Alternatively, in a given amount of space a larger matrix may be stored.

Consider the following 3 by 3 matrix. Not all the cross-products are computed, but only C(1, 1), C(2, 1), C(2, 2), C(3, 1), C(3, 2), C(3, 3). Now, let

$$\text{R(1)} = \text{C(1, 1)}$$

$$\text{R(2)} = \text{C(2, 1)}$$

$$\text{R(3)} = \text{C(2, 2)}$$

$$\text{R(4)} = \text{C(3, 1)}$$

$$\text{R(5)} = \text{C(3, 2)}$$

$$\text{R(6)} = \text{C(3, 3)}$$

and, in general, for square matrices, of any order,

$$\text{R(K)} = \text{C(I, J)}.$$

In this statement,  $K = (I*(I - 1))/2 + J$ . K is always an integer, since I and I - 1 are consecutive numbers, one of which must always be even. If J exceeds I, these two index values must be reversed in the above formula.

The coding is as follows:

$$\text{DO } 500 \text{ I} = 1, \text{ M}$$

$$\text{DO } 500 \text{ J} = 1, \text{ I}$$

$$K = (I*(I - 1))/2 + J$$

$$500 \text{ R}(K) = \text{R}(K) + \text{X}(I) * \text{X}(J)$$

Assuming that the sums corresponding to each variable have been stored in a column vector SX, and that the number of cases is S, then the following coding for the computation of means and standard deviations will appear somewhere in the program:

```
DO 600 I = 1, M
```

$$K = (I*(I + 1))/2$$

$$\text{AV}(I) = \text{SX}(I)/S$$

$$\text{VAR} = \text{R}(K)/S - \text{AV}(I) **2$$

$$\text{SD}(I) = 0$$

```
IF (VAR) 600, 600, 580
```

$$580 \text{ SD}(I) = \text{SQRT}(\text{VAR})$$

```
600 CONTINUE
```

Following this routine, the correlation coefficients may be computed:

```
DO 700 I = 1, M
```

```
DO 700 J = 1, I
```

$$K = (I*(I - 1))/2 + J$$

$$\text{DEN} = \text{SD}(I) * \text{SD}(J)$$

$$\text{R}(K) = 0$$

```
IF (DEN) 700, 700, 680
```

$$680 \text{ R}(K) = (\text{R}(K)/S - \text{AV}(I) * \text{AV}(J))/\text{DEN}$$

```
700 CONTINUE
```

The print-out routine must also provide for single-subscript notation. Also, if the correlation matrix is very large, it will be necessary to partition it both horizontally and vertically for printing. It is generally convenient to print 16 columns and 48 rows of correlations per page. The following routine will number the rows and columns, partition the square matrix, and perform the printing:

```
DO 800 II = 1, M, 16
```

$$\text{JJ} = \text{II} - 1 + 16$$

```
IF (M - JJ) 710, 720, 720
```

```

710 JJ = M
720 JK = JJ - II + 1
DO 800 KK = 1, M, 48
PRINT 19
PRINT 23, (I, I = II, JJ)
PRINT 18
LL = KK - 1 + 48
IF (M - LL) 730, 740, 740
730 LL = M
740 DO 800 I = KK, LL
DO 770 J = II, JJ
III = J - II + 1
IF (I - J) 760, 750, 750
750 K = (I* (I - 1))/2 + J
GO TO 770
760 K = (J* (J - 1))/2 + I
770 Q (III) = R(K)
800 PRINT 24, I, (Q(III), III = 1, JK)
18 FORMAT ('0')
19 FORMAT ('1')
23 FORMAT (18X, 16 (4X, 13))
24 FORMAT (11X, 14, 3X, 16F7.3)
DIMENSION Q(16)
DIMENSION AV(220), SD(220), SX(220), X(220)
DIMENSION R(24310)

```

It is a simple matter to expand this routine to include alphabetic titles for the rows and columns.

The above problem is dimensioned for 220 variables. The dimension for R equals  $(220 \times 221)/2$ . This can be increased or decreased as occasion demands and computer capacity permits.

Finally, it may be noted that while the assignments statements for K

add additional instructions to the program, the program with single subscripts compiles and executes faster than a program with double subscripts.

### *Effect of Single Subscripting on the Size of the Matrix*

As an illustration of the effect of single subscripting on the size of the maximum matrix, suppose that 24310 words are available for the cross-product matrix. If the double subscript notation is used, the maximum matrix that can be calculated is 155 by 155. If single subscripts are used, the maximum size matrix is increased to 220 by 220.

### *Availability*

The segments provided herein can be readily combined into the user's program to suit his requirements. However, a source listing, sample problem and documentation are available at cost (two dollars) from the author at New York University, School of Education, 32 Washington Place, New York, New York 10003.





## A COMPUTER PROGRAM TO CREATE A POPULATION WITH ANY DESIRED CENTROID AND COVARIANCE MATRIX

JOHN D. MORRIS

Rehabilitation Research Institute  
University of Florida

A computer program written in FORTRAN IV is presented which will create a population of desired size with marginally normal score vectors manifesting any desired centroid and covariance matrix. Uses and documentation are provided.

THE most usual and traditional direction for the statistical analysis of data has been that of reduction. In consonance with scientific parsimony, relatively large masses of data are quantitatively reduced to much simpler indices. There may be, however, occasions when the researcher may wish to proceed in the reverse direction of expansion. More specifically, since almost all parametric properties (excluding higher order moments) of a multivariate data set are subsumed within the centroid (vectors of means) and covariance matrix (or equivalently in the correlation matrix and variances), occasions may arise when the methodologist and/or the practical behavioral science researcher may wish to create a population with a desired number of score vectors which are marginally normal and manifest a desired centroid and covariance matrix. The purpose of this paper is to describe a computer program that will generate a population with these very characteristics.

### *Three Occasions for Generation of a Population with Specific Characteristics*

Three occasions are outlined for which the creation of such a population might be desired.

Copyright © 1975 by Frederic Kuder

### *Instructional Uses*

In the multivariate statistics or measurement classroom, the instructor could generate data exhibiting whatever parametric characteristics he might wish and could assign these to groups of students for analysis and interpretation. These data sets could include problems of easy interpretability (of the level of Cattell's "plasmode," 1966, p. 223), and relatively difficult interpretive problems. Possibilities for individualized instruction of students at different points along a graduated interpretive difficulty continuum also appear plausible.

### *Need to Accept Correlation Matrix as Input Data*

Most researchers using multivariate techniques have encountered the problems associated with a package of programs that will not accept a correlation matrix as input data. An example might be that a researcher would desire to use the missing data options of *Statistical Package for the Social Sciences* (Nie, Bent, and Hull, 1970) to create a correlation matrix and then would wish to complete a canonical correlation analysis. Since SPSS does not have a canonical correlation program and since the other most popular packages that do have a canonical correlation program may not accept a correlation matrix as raw data (BMD 09M, Dixon, 1973; *Statistical Analysis System*, Barr and Goodnight, 1972), the researcher may be at least temporarily stalled. The suggestion is that a sample manifesting the exact covariance matrix and centroid of the sample of interest could be created and these score vectors could be used as raw data. The results of such a run would be identical to using the original score vectors as input. This same problem would occur anytime a researcher wishes to do statistical analyses from a correlation or covariance matrix from the literature or elsewhere and finds that the common statistical packages will not accept the reduced input (Kerlinger and Pedhazur, 1973, p. 343).

### *Use in Monte Carlo Studies*

The last use is in Monte Carlo studies in which the methodological researcher wishes to draw samples from a population of score vectors with a desired centroid and covariance matrix.

### *Related Efforts*

Collier, Baker, Mandeville, and Hayes (1967) described a method for creating a population which manifests "approximately" a desired

covariance matrix. However, this method gives only an approximate solution. There apparently has been no information provided regarding the method's relative accuracy. Moreover, the degree to which the covariance matrix of the created sample fits that desired is contingent upon the size of the sample. Further literature search produced no documentation of any other proposed methods or computer programs.

### *Computational Procedure*

A population with the desired number of approximately random normal deviate score vectors is created by the Box and Muller (1958) method as modified by Marsaglia and Bray (1964). Each score vector has a number of components equal to the number of variables desired in the covariance matrix. Small intercorrelations between the random normal deviate variables are eliminated by triangularly decomposing (Guertin and Bailey, 1970) the random normal deviate intercorrelation matrix and by calculating factor scores which are thus standardized and independent. Each of these vectors is then premultiplied by the triangularly decomposed covariance matrix desired. The vector of desired means is then added to each score vector. This procedure results in the population with the desired number of score vectors with the desired centroid and covariance matrix, and by the central limit theorem the score vectors should be marginally normal. This normality has been consistently empirically verified with tests of properties of the resulting distributions (McNemar, 1962).

### *Program Input and Output*

An eight digit random seed number must be supplied to the program in the first eight columns of the first input card. On the second card the number of score vectors and the variables desired are entered in five column fields. In addition a "1" is punched if the special case of independent score vectors is desired. This option allows the program to skip the decomposition of a desired diagonal covariance matrix. The third card contains the variable format necessary for reading the input centroids and covariance matrix. If independent score vectors were chosen, one must then merely enter a vector of variable variances desired; otherwise, the entire variable covariance matrix must be entered.

The score vectors desired are punched, and the covariance matrix is calculated and compared with that input. The resulting difference matrix of errors is then printed. The errors are never larger than .01%;

however, if more precision is desired, double precision arithmetic could be easily instituted.

### Availability

The program with complete documentation, including a sample card set up and output are available upon request from the author.

### REFERENCES

- Barr, J. A. and Goodnight, J. H. *A users guide to the statistical analysis system*. Raleigh, North Carolina: Student Supply Stores, North Carolina State University, 1972.
- Box, G. E. P. and Muller, M. E. A note on the generation of random normal deviates. *Annals of Mathematical Statistics*, 1958, 28, 610-611.
- Cattell, R. B. (Ed.) *Handbook of multivariate experimental psychology*. Chicago: Rand-McNally, 1966.
- Collier, R. O., Baker, F. B., Mandeville, G. K., and Hayes, T. F. Estimates of test size for several test procedures based on conventional variance ratios in the repeated measures design. *Psychometrika*, 1967, 32, 339-353.
- Dixon, W. J. (BMD09M) In *Biomedical computer programs*. Los Angeles: University of California Press, 1973.
- Guertin, W. H. and Bailey, J. P. *Introduction to modern factor analysis*. Ann Arbor: Edwards Brothers, 1970.
- Kerlinger, F. N. and Pedhazur, E. J. *Multiple regression in behavioral research*. New York: Holt, Rinehart and Winston, 1973.
- McNemar, Q. *Psychological statistics*, New York: Wiley, 1962.
- Marsaglia, G. and Bray, T. A. A convenient method for generating normal variables. *SIAM Review*, 1964, 6, 260-264.
- Nie, N. H., Bent, D. H., and Hull, C. H. *Statistical package for the social sciences*. New York: McGraw-Hill, 1970.

## CORALL, A FORTRAN IV PROGRAM FOR CORRELATION MEASURES

JAN VEGELIUS  
University of Uppsala

The program can compute a great number of different correlation and other statistical measures. The user is free to select among the measures and also among the variables that are read by the program. When a particular set of variables has been treated in the prescribed way, a new set may follow together with new measure definitions.

In Vegelius (1973) a great number of different correlation measures is described. Most of them (together with various other elementary statistical measures) are available in a FORTRAN IV program called CORALL (Vegelius, 1974), constructed in a way similar to the SIEGEL program (Vegelius, 1971).

### *Contents*

This program makes available 20 correlation measures, 3 measures of central tendency, 12 other elementary measures, 4 instructions for file treatment and 7 other instructions, e.g., transposition of the data matrix, comments for the output, dichotomization of the data, and treatments of missing data. The tetrachoric correlation coefficient will be computed from the first terms in the power series. The number of terms may be chosen by the user. Up to 36 terms are possible.

### *Input*

The values are normally to be read variable-wise, but it is also possible to read person-wise and to transpose the data-matrix. Both variable-format input and format-free input can be used. If one card



(containing 80 columns) is not enough when punching the format, up to 4 extra cards may be used. When the data are read from cards or from tape, the measure instructions may follow. It is then possible to select any of the samples and any of the measures. After one measure has been chosen, a new one may follow. There is no upper limit to the number of measure instructions following the data input.

*Example:* If the data have been read, and the means, the standard deviations, the indexes of skewnesses, the indexes of kurtosis of all the variables, and the product moment correlation coefficients between each variable pair are desired, the card order may be (each card beginning in the first column):

MEAN  
STDEV  
SKEWNESS  
KURTOSIS  
PEARSON

### *Restrictions*

The program has been written in FORTRAN IV for use on an IBM 360/370. As all measures are available in the memory at the same time, the program works rather fast. This kind of availability means, however, that it is not possible to treat a great amount of data at the same time. In the three existing versions, the upper limit to the number of data is 2000 (156 K), 9000 (208 K), and 25,000 (416 K).

### *Availability*

A description of the program may be obtained from Jan Vegelius, Department of Psychology, Svartbäcksgatan 10, S 753 20 Uppsala, SWEDEN.

## REFERENCES

- Vegelius, J. SIEGEL, a FORTRAN program for nonparametrical methods. Uppsala: University of Uppsala, Department of Psychology, 1971 (*Report 109*).
- Vegelius, J. Correlation coefficients as scalar products in Euclidean spaces. Uppsala: University of Uppsala, Department of Psychology, 1973 (*Report 145*).
- Vegelius, J. CORALL, a FORTRAN program for correlation measures. Uppsala: University of Uppsala, Department of Psychology, 1974 (*Report 147*).

## SIEGEL, A FORTRAN IV PROGRAM FOR NONPARAMETRICAL METHODS

JAN VEGELIUS  
University of Uppsala

The program can perform any one of those nonparametric statistical methods that is mentioned in Sidney Siegel's classical work on the subject. The user is free to select among the methods and also among the samples that are read by the program. When a particular set of samples has been treated in the prescribed way, a new set may follow together with new method definitions.

In Siegel (1956), a large variety of nonparametric statistical methods is described. All of them (together with a method for multivariate information analysis, described by Attneave (1959)) are available in a FORTRAN IV program called SIEGEL (Vegelius, 1971, 1974a, 1974b).

### *Contents*

This program makes available 4 one-sample methods, 16 two-sample methods, and 8 methods for more than two samples. For file treatments, there exist 4 instructions. Finally, six extra instructions are available, e.g., transposition of the data matrix (if it is square), comments to be written in the output, and treatments of missing data.

### *Input*

The values are normally to be read sample-wise, and it is possible to have different sizes of the samples. Both variable-format input and format-free input can be used. If one card (containing 80 columns) is not enough when punching the format, up to 4 extra cards may be

used. When the data are read from cards or from tape, the method instructions may follow. It is then possible to select any of the samples and any of the methods. After one method has been employed, a new one may follow. There is no upper limit to the number of method instructions following the data input.

*Example:* If two independent samples have been read and Mann-Whitney's U-test, Wald-Wolfowitz' run-test, and Moses' test of extreme reactions are desired, the card order may be (each card beginning in the first column):

UTEST

WALD

MOSES

\*

### *Output*

For most of the methods it is possible to vary the amount of output. If more than 2 samples have been selected for the two-sample tests, the output will be the elements below the main diagonal in a square matrix, where the elements have been obtained by a comparison between all possible sample pairs. The values used are probabilities or correlation coefficients. For further details, see Vegelius (1971).

### *Restrictions*

The program has been written in FORTRAN IV for use on an IBM 360/370. As all methods are available in the memory at the same time, the program works rather fast. This kind of availability means, however, that it is not possible to treat a great amount of data at the same time. In the two existing versions, the upper limit for the number of data is 2000 (156 K), and 6000 (208 K), respectively.

### *Availability*

A description of the program may be obtained from Jan Vegelius, Department of Psychology, Svartbäcksgatan 10, S 753 20 Uppsala, SWEDEN.

## REFERENCES

- Attneave, F. *Applications of information theory to psychology*. New York: Henry Holt, 1959.
- Siegel, S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.

- Vegelius, J. SIEGEL, a FORTRAN program for nonparametrical methods. Uppsala: University of Uppsala, Department of Psychology, 1971 (*Report 109*).
- Vegelius, J. CORALL, a FORTRAN program for correlation measures. Uppsala: University of Uppsala, Department of Psychology, 1974a (*Report 147*).
- Vegelius, J. Changes in the new version of the SIEGEL program. Uppsala: University of Uppsala, Department of Psychology, 1974b (Mimeo).





## A COMPUTER PROGRAM FOR FISHER'S EXACT PROBABILITY TEST

D. L. TRITCHLER AND D. T. PEDRINI  
University of Nebraska at Omaha

The Fisher test is concerned with data of independent groups and a dichotomous criterion in a fourfold ( $2 \times 2$ ) contingency format. Typically, small samples ( $N \leq 30$ ) are considered. Some tables of critical values of Fisher's test for small samples are available to offset the tediousness of computation. Fisher's test requires determination of a combined probability, that is, the observed set added with all the more extreme (directional) sets. For large samples, an approximation method such as chi square is used. Considered in this paper are a mathematical discussion and algorithm, and a computer program for Fisher's test. The computer program for small and large ( $N$  of about 500) samples is available from the authors. Such technology, in most instances, makes approximation methods unnecessary.

A common statistical problem is the determination of the significance of a difference between two independent samples or groups for which the scores fall into two mutually exclusive classes. This problem is commonly depicted in the form of a fourfold or a  $2 \times 2$  contingency table. As a technique for analyzing such data, Fisher's exact test provides a precise probability. Since this test becomes impractical for hand computation in the case of all but the smallest samples, usually an approximation method such as chi square is chosen. Considered in this paper are a mathematical discussion and algorithm, and a computer program for Fisher's test.

### *Mathematical Discussion*

If the marginal totals of a given  $2 \times 2$  contingency table are considered fixed, the exact probability of a given set of frequencies, given

the null hypothesis of no difference in population proportions, is given by the hypergeometric distribution. In Fisher's test the null hypothesis requires a determination of the combined probability of the observed set of frequencies and all sets of frequencies more extreme than the observed (for example, see Blalock, 1972, pp. 287-291). Tritchler and Pedrini (1974) developed a mathematical algorithm which is computationally efficient and accurate.

### *The Computer Program*

Tritchler and Pedrini (1975) developed a FORTRAN computer program to implement the algorithm referred to above. The program is appropriate for sample sizes ranging from multiple zero cells to an  $N$  of about 500 (as analyzed on an IBM 360/65).

Of course, for small sample sizes, desk calculations may suffice. Some nonparametric texts (for example, Siegel, 1956, pp. 256-270) include tables of critical values of Fisher's test for small samples. But for large sample sizes and for exact probabilities, a computer program seems mandatory.

### *Input*

The data for a given problem are entered on the data card. Let  $a, b$  be the class frequencies for one group and let  $c, d$  be the respective class frequencies for the other group. Then the integers are entered on the card in the order  $a, b, c, d$ , separated by blanks. The input is free-format, that is, the integers may be anywhere on the card.

### *Output*

Output consists of the contingency table with margin totals, and Fisher's exact probability for the one-tailed case. A sample output is shown below:

OUTCOME				
G		+	-	
R	X	34	36	70
O	Y	444	185	629
U		478	221	699
P				

P (OCCURRENCE OF A DIFFERENCE THIS EXTREME OR GREATER) = 0.0002 NOTE—THIS IS A ONE-TAILED TEST

### *Capabilities and Limitations*

The algorithm used by the program is quite accurate, with a maximum relative error of  $(6a + 2c - 2)/10^{M-1}$  (where  $a$  and  $c$  are the fre-

quencies in the observed cells A and C, and M is the number of significant digits of a double precision number for the machine being used). Because of the large values involved in computing factorials, values are repeatedly scaled to prevent exponent overflow. If the procedure breaks down (contingent upon the sample size, the distribution among cells, and the computer used), the fact is noted in the output (Tritchler and Pedrini, 1975). Robertson (1960) and Gregory (1973) also have discussed computer programs for Fisher's test, although Robertson has not mentioned maximum sample size limitations, Gregory has defined an upper limit of  $N = 100$  for his program.

### *Availability*

A copy of this article and a program listing can be obtained from D. L. Tritchler, Computer Center, or D. T. Pedrini, Psychology Department, University of Nebraska at Omaha, Omaha, Nebraska 68132.

### REFERENCES

- Blalock, H. M. *Social statistics*. (2nd ed.) New York: McGraw-Hill, 1972.
- Gregory, R. J. A FORTRAN computer program for the Fisher exact probability test. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1973, 33, 697-700.
- Robertson, W. H. Programming Fisher's exact method of comparing two percentages. *Technometrics*, 1960, 2, 103-107.
- Siegel, S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.
- Tritchler, D. L. and Pedrini, D. T. *Computer technology, large samples, and Fisher's exact probability test*. Omaha Neb.: University of Nebraska at Omaha, 1974. (ERIC Document Reproduction Service No. ED 093 336).
- Tritchler, D. L. and Pedrini, D. T. *Fishept: A computer program for Fisher's exact probability test*. Unpublished manuscript, University of Nebraska at Omaha, 1975.



## A COMPUTER PROGRAM TO DETERMINE RELATIONS AMONG GENUINE DICHOTOMIES: THE PHI AND G STATISTICS

HOWARD CHAMBERLAIN AND DAVID D. VAN FLEET  
Texas A&M University

The Phi and G statistics for dichotomus variables are discussed and a Fortran program to compute them is described. Input is to be in card form, output may be printed, punched, or placed on magnetic tape. The punch or tape output is designed to be used as input for the BMD X72 factor analysis program.

IN some types of research the responses of the subject are in the form of true dichotomies; such as, yes-no, true-false or existence or absence of a given condition. Two statistics suggested for this situation are Phi and G.

Phi is the product moment coefficient between pairs of genuinely dichotomous variables. Except under certain conditions, it is not symetrically distributed about zero and does not range from  $-1.0$  to  $+1.0$  (Guilford, 1965). Additionally, Phi may have a sign indicating a clearly nonlogical direction of relation (Holley and Guilford, 1964). G is based upon the probability of agreement of response, is symetrically distributed, and ranges from  $-1.0$  to  $+1.0$  (Holley and Sjoberg, 1968). While Phi is sensitive to skewness of the marginal distributions, G is not (Holley and Eriksson, 1970).

The purpose of this paper is to present a computer program that will calculate the Phi and G coefficients. It is designed to take up to 400 observations and 150 variables or questions. However, the user may adjust the program to fit his requirements.

**Input:** The input must be in card form with responses punched as zeros and ones or as blanks and ones. The user controls the form of the input through a format card.



**Output:** The output will be a table of values in the form required as input to the BMD X72 factor analysis program. Through use of a control card any combination of printer, punch, or tape output may be selected. The control card also allows the user to suppress either the Phi or G table.

This Fortran program is available from Dr. Howard Chamberlain, Department of Management, Texas A&M University, College Station, Texas 77843.

## REFERENCES

- Guilford, J. P. *Fundamental statistics in psychology and education*. New York: McGraw-Hill, 1965.
- Holley, J. W. and Eriksson, U. A note on the effect of selective sampling procedures on the Phi coefficient. *Multivariate Behavioral Research*, 1970, 5, 117-123.
- Holley, J. W. and Guilford, J. P. A note on the G index of agreement. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1964, 24, 759-753.
- Holley, J. W. and Sjöberg, L. Some characteristics of the G index of agreement. *Multivariate Behavioral Research*, 1968, 3, 107-114.

## EDUC1 LIBRARY: A DESCRIPTION OF FORTRAN IV COMPUTER PROGRAMS FOR THE IBM SYSTEMS 3/10

RICHARD B. BALDAUF, JR.

Department of Education, American Samoa<sup>1,2</sup>

A library of 20 FORTRAN computer programs has been compiled, modified, and edited to provide in a single source a series of test scoring, data reduction, and evaluation programs for educators having access to small business-oriented computers. Summary details are provided for each program.

A growing number of school systems and users who have access to small core business oriented computers with limited FORTRAN compilers are finding that there is no single source of suitable programs to solve the educational problems of test scoring, data reduction, and program evaluation. The EDUC1 library, which was written to meet these needs, is a selection of twenty computer programs which have been edited and rewritten to conform to the FORTRAN compiler supported by the IBM Systems 3/10 and to have a maximum size of 32K. (International Business Machines, 1972). When reducing program size, the writer made every effort to retain or add desirable features and to standardize input procedures so as to make the programs relatively easy to use for educational program managers and evaluators unfamiliar with computer programming.

### *Modifications*

Modification in the programs were of two types. First, programs were altered to conform to the Systems 3/10 FORTRAN compiler.

<sup>1</sup> The programmer is indebted to Robert V. Bloedon for his initial assistance and support, without which this project would not have been undertaken, and to Sili Atuatasi, Assistant Director for Research and Development, and Mere T. Betham, Director of Education, who allocated the resources necessary for its completion.

<sup>2</sup> Now with the Faculty of Education, James Cook University, Australia.  
Copyright © 1975 by Frederic Kuder

TABLE 1  
*EDUCI: Program Identification, Original Source, and Major Statistical Output*

Program Name	Program Description	Original Source	Mean	S.D.	Correlations	Distribution	Output Significance Test	Factor Loadings	Pct. Variance	Item Analysis
ITEM	Test scoring	Oosterhof & Kocher (1972)	X	X	X	X				X
TESTAT	Scale and test scoring	Veldman (1967)	X	X	X	X				X
DISTAT	Distribution statistics	Veldman (1967)	X	X	X	X				
CORREL	Correlational analysis	Bloedon (1974)	X	X	X					
MDCORS	Missing data correlation	Veldman (1967)	X	X	X					
TDCORS	Transpose data correlation	Veldman (1967)	X	X	X					
FACTOR	Principal/Varimax analysis	Veldman (1967); Bloedon (1974)	X	X	X			X	X	
FSORE	Factor scores for subjects	Veldman (1967)								
RELATE	Comparison factor matrixes	Veldman (1967)			X			X	X	
VECTOR	Vector-powered factoring	Overall & Klett (1972)	X	X	X			X		
ANOVAR	Repeated measures ANOVA	Veldman (1967); Barker (1973)	X				X			
AVAR23	2 x 3 Factorial ANOVA	Veldman (1967); Barker (1973)	X				X			
STEPWZ	Stepwise regression	Barker (1973)	X	X	X		X		X	
LINMOD	Linear models, regression	Dixon (1970)	X	X	X		X		X	
		Ward & Jennings (1973)	X	X	X		X		X	
DISCRM	Discriminant analysis	Veldman (1967)	X		X		X		X	
CANONA	Canonical correlation	Veldman (1967); Cooley & Lohnes (1971)	X	X	X		X	X	X	
TABLE	Cross-tabulation	Bloedon (1974)				X				
CHICHI	Chi-square	Veldman (1967)				X				
LCLUST	Linear typal analysis	Overall & Klett (1972)	X		X			X		
WDLIST	Content analysis	Veldman (1967)				X				

TABLE 2  
EDUC1: Description of Programs' Major Characteristics

Program Name	Largest Size <sup>a</sup>	Number of Variables	Choices Levels Factors	Disk Input	Missing Data	Real Input	Disk <sup>b</sup> Storage Needed	Batch Jobs	Program Units <sup>c</sup>
ITEM	27.9K	90	5	YES	YES	NO <sup>d</sup>	YES(1)	YES	1
TESTAT	28.0K	210	9, 13	YES	YES	NO <sup>d</sup>	YES(1)	YES	2
DISTAT	28.2K	10	120	YES	YES	YES	YES(1)	YES	3
CORREL	27.0K	75	—	YES	NO	YES	NO	YES	1
MDCORS	27.0K	30	—	YES	YES	YES	NO	YES	1
TDCORS	28.9K	15	50	YES	NO	YES	NO	YES	1
FACTOR	28.4K	50	12	NO	NO	YES	NO	NO	3
FSCORE	16.6K	50	12	YES	NO	YES	NO	YES	1
RELATE	28.4K	50	12, 12	YES	NO	YES	NO	YES	1
VECTOR	27.4K	50	—	YES	NO	YES	YES(1)	YES	1
ANOVAR	29.5K	40	10, 4	YES	YES	YES	NO	YES	3
AVAR23	28.4K	40	6, 6, 6	YES	YES	YES	YES(1)	YES	3
STEPWZ	30.1K	25	—	YES	NO	YES	YES(2)	YES	6
LINMOD	27.6K	25	—	YES	NO	YES	YES(1)	YES	2
DISCRM	29.1K	20	9	YES	NO	YES	YES(1)	YES	5
CANONA	28.0K	22, 22	—	NO	NO	YES	YES(2)	NO	6
TABLE	18.3K	99	12	YES	YES	NO <sup>d</sup>	YES(1)	YES	1
CHICHI	26.2K	40	40	NO	NO	YES	NO	YES	2
LCLUST	27.4K	80	300	YES	NO	YES	YES(4)	NO	3
WDLIST	28.4K	12	1200	NO	YES	NO	NO	YES	2

<sup>a</sup> The largest core necessary to run the program is given.<sup>b</sup> Number of disk storage files needed is given in parentheses.<sup>c</sup> Number of programs, subroutines and functions making up the program<sup>d</sup> Only integer data are acceptable.

For example, because variable formats are not supported, the programs were written with a choice of four standard formats. Many subroutines had to be eliminated since the compiler does not allow them to be dynamically dimensioned. Others were deleted because subroutines can not call other user supplied subprograms, nor contain any disk file operations (IBM, 1972). Second, program size was reduced by limiting the number of variables, and subroutines were changed or eliminated if they required separate core storage above that available in the main program. Programs were generally restricted in size to about 28K to avoid overlays which were found to increase processing time considerably. Many of the programs require one or more storage files to operate.

These changes have resulted in a series of programs which contain few overlapping components. Although this simplification is far from ideal from the programmer's point of view, it has resulted in programs which can handle relatively large amounts of data with only restricted core and at a reasonable cost. Users with computers having less than 32K could use many of the programs in the library by reducing the dimensioned size of the variables.

### *Descriptions*

Table 1 lists the programs' names, provides a brief description of the analyses computed, and gives the original program source. The EDUC1 users Manual (Baldauf, 1974) furnishes a detailed description of how to set up and to operate each program along with a brief summary and an example. Users of these programs will probably want to consult the original sources for each to obtain descriptions of computational details, program functions, and uses.

### *Characteristics*

Table 2 lists for each program details concerning program size, capabilities, and limitations. In addition to the programs listed, the library provides a series of simple utility routines.

### *Availability*

Copies of the EDUC1 users manual, sample test data, program source listings, and program source decks are available at cost from the Supervisor of Testing, Department of Education, Pago Pago, American Samoa.

## REFERENCES

- Baldauf, R. B., Jr. *EDUC1 users manual*. Pago Pago: Department of Education, 1974.



- Barker, D. G. Factorial analysis of variance with Scheffe's test. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1973, 33, 741. (a)
- Barker, D. G. Groups by trials analysis of variance with Scheffe's test. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1973, 33, 743. (b)
- Bloedon, R. V. *EDSTAT: A source statement listing*. Honolulu: Education Research and Development Center, 1974.
- Cooley, W. W. and Lohnes, P. R. *Multivariate data analysis*. New York: Wiley, 1971.
- Dixon, W. J. (Ed.) *BMD: Biomedical computer programs*. Berkeley: University of California Press, 1970, 233-257d.
- International Business Machines, *IBM Systems/3 disk FORTRAN reference manual*. Rochester Minn: IBM, 1972.
- Oosterhof, A. C. and Kocher, A. T. An item analysis program which provides feedback to individual students. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1972, 32, 799-800.
- Overall, J. E. and Klett, J. C. *Applied multivariate analysis*. New York: McGraw-Hill, 1972.
- Veldman, D. *FORTTRAN programming for the behavioral sciences*. New York: Holt, Rinehart & Winston, 1967.
- Ward, J. H. and Jennings, E. *Introduction to linear models*. Englewood Cliffs, N. J.: Prentice-Hall, 1973.



## FORTRAN IV PROGRAM TO DETERMINE THE PROPER SEQUENCE OF RECORDS IN A DATAFILE<sup>1</sup>

MICHAEL P. JONES AND ROLAND K. YOSHIDA<sup>2</sup>

Neuropsychiatric Institute-Pacific State Research Group  
Pomona, California

This FORTRAN IV program executes an essential editing procedure which determines whether a datafile contains an equal number of records (cards) per case which are also in the intended sequential order. The program which requires very little background in computer programming is designed primarily for the user of packaged statistical procedures.

In the preparation of datafiles for analysis, various types of errors arise which violate the requirement of a sequentially ordered file with an equal number of records (cards) per case. The purpose of this paper is to describe a program that is designed to identify two common difficulties: (a) the existence of too few or too many records for a case and (b) the improper sequencing of cards. This program is especially useful for researchers who have relatively little background in computer languages and who rely primarily upon packaged statistical programs such as SPSS (Nie, Hull, Jenkins, Steinbrenner, and Bent, 1975) and BMD (Dixon, 1973).

### *Input Information*

There is no limit on the number of records in the datafile of interest; input source may be from card image, disk, or tape. Case and card

<sup>1</sup> This report was funded in part by a U.S. Office of Education, Bureau of the Handicapped Project OEG 0-73-5263. The opinions expressed herein do not necessarily reflect the position or policy of the U.S. Office of Education, and no official endorsement by the U.S. Office of Education should be inferred.

<sup>2</sup> Also at the University of Southern California.

numbers may be located in any column of a record, and two inclusive ranges of card numbers may be searched such as 1-9 and 15-23.

The routine, which is written in FORTRAN IV, requires these control cards:

1. **PARAMETER CARD:** (mandatory) Format (2A4, A2, I2, 1X, 4(I2, 1X))

Col 1-9 Code **PARAMETER**

Col 11-12 (mandatory) Identifies the input source. Any unit number (right justified) from 1-99 except 6 and 7 is valid (5-card reader only).

Col 14-15 (mandatory) Identifies the beginning card number of the first sequence (right justified).

Col 17-18 (mandatory) Identifies the ending card number of first sequence (right justified).

Col 20-21 (optional) Identifies the beginning card number of the second sequence (right justified).

Col 23-24 (optional) Identifies the ending card number of second sequence (right justified).

2. **FORMAT CARD:** (mandatory) Format (2A3, 18A4)

Col 1-6 Code **FORMAT**

Col 7-80 Code **FORTRAN** input format for case number and card number.

The data cards follow the format card. If the input source is disk or tape, insert the proper FT statement corresponding to the unit coded on the **PARAMETER** card.

### *Limitations*

The following limitations apply to the card order program:

1. Case IDs, which must be numerical, may not exceed 9 characters in length.
2. Card numbers must range between 1 and 99.
3. Case numbers are not searched for sequence, duplication, or inclusion in the datafile.

### *Output*

For each datafile, the output specifies the card numbers of each case which violates the sequence and the range of card number values provided on the parameter card and gives the total number of cards for the datafile of interest.

### *Availability of Program*

A listing and write-up of the card order program along with sample input and output can be obtained by writing to Roland K. Yoshida,

Neuropsychiatric Institute-Pacific State Research Group, P. O. Box 100-R, Pomona, California 91766.

### REFERENCES

- Dixon, W. J. (Ed.) *BMD: Biomedical computer programs* (3rd Ed.). Berkeley and Los Angeles: University of California Press, 1973.
- Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., and Bent, D. H. *SPSS: Statistical package for the social sciences*. (2nd Ed.). New York: McGraw-Hill, 1975.





## A FORTRAN PROGRAM FOR ANALYZING THE RESULTS OF FLANDERS' INTERACTION MATRIX: AN UPDATED VERSION

VINCENT RACIOPPO AND PIETRO J. PASCALE  
Youngstown State University

GAVIN DOUGHTY, JR.  
Tarkio College

This paper presents a revised and updated version of a FORTRAN program which computes all indices used in the Flanders' Interaction Matrix. The new program has added another form of data input which simplifies data entry. The new version also has the capability of interactive terminal use.

RACIOPPO and Pascale (1974) developed a FORTRAN program for analyzing the outcomes of Flanders' Interaction Matrix. The original program has been updated and improved. Several improvements such as the capability for use of the program on keyboard terminals have been developed. The revision feature of this program which supports keyboard terminal use is the new manner of data entry. Data entry for the original program required data to be in matrix form. In other words, the user had to hand tally the data into a ten by ten matrix.

### *Purpose*

The purpose of this paper is to make known a new program that provides for data entry in the form of Flanders' category numbers in direct correspondence to the observation. The program in effect builds the ten by ten matrix.

### *Program Characteristics*

What the observer records is the input which opens the possibility for putting a portable keyboard terminal into a classroom and

recording the categories directly onto the keyboard. There is virtual instantaneous evaluation of both the configuration of the ten by ten matrix and the various indices. The revised program also defines on comment cards all the computed indices.

### *Availability*

Those who have requested the initial program will receive the updated version. Copies of the manuscript, a complete documentation, a listing, and the program punched on cards can be obtained from Dr. Pietro J. Pascale, Youngstown State University, Foundations of Education, Youngstown, Ohio 44555.

### REFERENCE

- Racioppo, V. and Pascale, P. J. A FORTRAN program for analyzing the results of Flanders' Interactional Matrix. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1974, 34, 711-712.

## A COMPUTER PROGRAM FOR CALCULATING AN INDEX OF INTEROBSERVER RELIABILITY FROM TIMESERIES DATA

BILLY W. THORNTON

Boys Town, Nebraska

FRANK L. CROSKEY

Shawnee Mission School District

The purpose of this paper is to describe a method of indexing interobserver reliability from data relating to mutually exclusive nominal categories. The computer program to be described computes estimates of reliability from nominal data collected in a time interval manner. A specified critical time interval, the order of the responses, and the number of responses are important factors in computing the amount of agreement between judges.

THE purpose of this paper is to describe a method of indexing interobserver reliability from data relating to mutually exclusive nominal categories. The computer program to be described computes estimates of reliability from nominal data collected in a time interval manner. A specified critical time interval, the order of the responses, and the number of responses are important factors in computing the amount of agreement between judges.

### *Background Information*

The computational procedures used in this program parallel steps outlined by Scott (1955). The formula for Scott's Coefficient of Interobserver Reliability is:

$$\pi = \frac{P_o - P_e}{1 - P_e}$$

$P_o$  (observed percentage of agreement) represents the percentage of judgments on which two observers agree when coding the same data independently;  $P_e$  is the percentage of agreement to be expected by chance. Thus,  $\pi$  is the ratio of the actual difference between obtained and chance agreement. Procedures for the hand calculation of  $\pi$  may be found in Flanders (1967, pp. 158-166). Ordinarily, the value of  $P_o$  is computed by considering number of agreements across categories for the complete set of data for two judges. This program computes  $P_o$  for two judges by summing the number of agreements within each time interval across time. Thus,  $P_o$  for judge  $j_1$  and judge  $j_2$  is expressed as follows:

$$P_o = \frac{\text{number of agreements}}{\text{number of total responses}}$$

The number of agreements for two judges is defined to be the sum over time intervals and categories of the minimum frequency of common observations per cell. For example, in a given time unit and for a specific category, if judge ( $j_1$ ) indicated three observations of the event and judge ( $j_2$ ) indicated four observations, then the two judges agree three times and disagree one time. It is assumed that if two judges indicate an event occurred, then there is agreement. However, if one judge indicates an event occurred and the other did not, then there is a disagreement. Therefore, the number of disagreements for two judges is the sum of the absolute value of the differences between the number of responses in corresponding cells. The total number of responses is the sum of the number of disagreements and number of disagreements.

### *General Description of the Computer Program*

The program is an interactive time-sharing FORTRAN program. The program asks a series of questions, the answers to which specify the number of judges, the time interval, the title of the run, and the name of the input data file. The input for the program consists of a three dimensional (judges, time, categories) data matrix; the entries of which are the number of observations for each cell.

Output consists of the number of agreements and disagreements between each pair of judges and interobserver reliability coefficients for each pair of judges. The coefficients are output in matrix form. The computer program allows the user the option of combining sets of consecutive intervals for which new coefficients are computed. The process can be repeated as desired. If the interval length equals the total time segment, the reliability coefficients between judges will attain maximum values. Thus, by comparing the magnitude of reliability coefficients from small intervals with reliability coefficients from larger



intervals, characteristics of the data are revealed. In interpreting the reliability coefficient, the user should keep in mind the possibility of a time lag between judges.

### *Program Availability*

The computer program which is written in time-sharing FORTRAN, operates on a Honeywell 635. A listing of the program is available on request from B. W. Thornton.

### REFERENCES

- Flanders, N. A. The problems of observer training and reliability. In E. J. Amidon and J. B. Hough (Eds.) *Interaction Analysis: Theory, Research, and Application*. Reading, Mass. Addison-Wesley, 1967.
- Scott, E. A., Reliability of content analysis: The case for nominal scale coding. *Public Opinion Quarterly*, 1955, 19, 321-325.



## BOOK REVIEWS

C. Mauritz Lindvall and Anthony J. Nitko. *Measuring Pupil Achievement and Aptitude*. (2nd. ed.) New York: Harcourt Brace Jovanovich, 1975. Pp. xi + 237. \$3.95 (paperback)

The first chapter of this book specifies the areas of pupil evaluation: (1) achievement, (2) aptitude, (3) interest, and (4) personality. The procedures of assessing these areas are then briefly outlined. The chapter concludes with several references relevant to the historical development of testing and evaluation and to more extended discussion of interest inventories and personality tests.

Chapter 2 is concerned with the planning of instruction and of evaluation. The emphasis on instructional objectives, their organization and sequencing. There is brief, but adequate, discussion of such characteristics of tests as validity, reliability, objectivity, and comprehensiveness, but no mention of machine scoring or analysis. In Chapter 3 different types of tests are related to different objectives as defined in *Taxonomy of Educational Objectives* by Benjamin Bloom and others (including this reviewer). Chapter 4 explains the construction of teacher-made tests. It is most adequate with reference to rules for writing different types of objective items, but least adequate with reference to the production of thought-provoking exercises.

Chapter 5 is devoted to the interpretation of test scores. It includes brief discussion of the difference between criterion-referenced and norm-referenced testing and brief, but adequate, explanations of percentile rank, mean, standard deviation, standard scores, normal distributions, and stanine scores. Table 5.4 presents excellent comparison of various kinds of norm-referenced scores. In Chapter 6 there is excellent elementary explanation of coefficients of correlation and of the kinds of test validity and of the means of assessing test reliability.

Chapter 7 explains the process of constructing and standardizing a norm-referenced achievement test and describes several widely used achievement test batteries including the *California Achievement Tests*, the *Iowa Tests of Educational Development*, the *Metropolitan Achievement Tests* and the *Stanford Achievement Tests*. Similarly Chapter 8 is concerned with scholastic aptitude—the Stanford-Binet and such group tests as the Otis-Lennon.

Chapter 9 deals with testing and evaluation in the individualizing of

instruction while Chapter 10 describes the planning and implementing of a comprehensive evaluation program. Some attention is given to the use of rating scales.

With some supplementation, this text would be an excellent basis of instruction for an introductory course in educational and psychological measurement.

MAX D. ENGELHART

Jerome M. Sattler. *Assessment of Children's Intelligence*. (Rev. reprint) Philadelphia: W. B. Saunders, 1974. Pp. xxii + 591. \$14.95.

The author lists three goals of this book: (1) to assist students with the process of psychological evaluation; (2) to bring out the findings and insights of many pioneer clinicians, educators, and investigators; (3) to summarize and integrate the findings of studies concerned with individual intelligence tests and variables in the testing situation. The major assumption underlying these goals is that individually-administered intelligence tests can yield much more than just an IQ. Furthermore, it is maintained that group-administered tests are less sensitive than individual tests to the cognitive and conative features of personality.

This large book possesses features of both a textbook and a source book, and this reviewer readily recommends it for graduate courses on intelligence testing. The book is well-written and clear, but the reader will need some knowledge of tests and measurements, developmental psychology, and abnormal psychology in order to derive the greatest benefit. An attractive feature for instructions is a manual of multiple choice questions. These questions, however, should be viewed as only a supplement to a more thorough evaluation of the students from observations of their performance in test situations and the quality of case reports.

The 27 chapters comprising the book are grouped into six sections: 1. Introduction and General Considerations; 2. Administering Individual Intelligence Tests; 3. Stanford-Binet Intelligence Scale; 4. WISC, WPPSI, and Other Tests; 5. Diagnostic Applications; 6. Psychological Reports and Consultation. The core of the book consists of detailed descriptions of the development, administration, and interpretation of the Stanford-Binet, WISC, and WPPSI. These chapters (8-17) by themselves constitute almost an entire course on intelligence testing of children. The reviewer found, however, that some of the most informative and useful material appears in Sections 2, 5, and 6. Admittedly, Section 5 (Diagnostic Implications) is a bit disappointing if one expects it to present clear-cut methods of diagnosing and prescribing for various disorders and exceptionalities. Such straightforward methods, of course, do not exist, and Jerome Sattler is no Pollyanna. Rather he is an empiricist and compiler who recognizes the limitations of intelligence tests for diagnostic and prescriptive pur-

poses. Consequently, hundreds of empirical investigations are cited in this book, but students and practicing psychologists who entertain unrealistic expectations about the diagnostic abilities of individual intelligence tests may "come out by that same door wherein (they) went."

In short, the twenty-seven chapters make up a comprehensive handbook on intelligence testing of children. Complementing these chapters are six appendixes: A. Supplementary WISC Scoring Criteria; B. List of Validity and Reliability Studies for the Stanford-Binet, WISC, WPPSI, PPVT, Quick Test, Leiter, and Slosson; C. Miscellaneous Tables; D. Stanford-Binet Intelligence Scale, Form L-M, 1972 Norms; E. Wechsler Intelligence Scale for Children-Revised (WISC-R); F. WISC-R Tables. The last three appendixes were added in the revision to accommodate the 1972 Stanford-Binet norms and the WISC-R. Anyone planning to use the WISC-R would be well advised to examine Appendixes E and F of this book before proceeding.

Unfortunately, not all individual intelligence testers possess the interpersonal sensitivity and statistical-technical expertise needed to draw sound diagnostic conclusions from test responses. To be sure, Jensen and Shockley have had some negative effects on the public image of intelligence testing. But this reviewer is convinced that vociferous critics of intelligence testing find ample ammunition in the errors of poorly-trained, and perhaps poorly-endowed, psychodiagnosticians. Many of these itinerant Binet-testers and WISC-testers might benefit from a serious study of Sattler's book. He has performed a useful scholarly service in collecting under one cover a great deal of material concerned with intelligence testing of children.

LEWIS R. AIKEN, JR.

Warren W. Willingham. *College Placement and Exemption*. New York: College Entrance Examination Board, 1974. Pp. xv + 272. \$6.95 and \$4.95 (paperback).

Bringing order out of chaos, or at least out of great diversity, is no small task, but this was one of the problems confronting the author of this text. American education is known for its diversity and these differences are clearly illustrated in the wide range of prevalent policies and practices in college placement and exemption. As stated by the author the primary purposes of this text are:

1. to develop a framework that would include the most important types of placement and exemption and closely related models and to help clarify the relationship among them.
  2. to describe the educational rationale and technical characteristics of these models.
  3. to review fairly thoroughly the relevant research literature.
- By accomplishing these purposes the primary aim "was to encourage on individual campuses more systemic analysis of the objectives and outcomes of these various models of sorting students into



alternate educational treatments." Even though, as expressed by the author, his intent was not to produce a "how to" handbook, this text is intended for responsible practitioners such as college administrators and faculty as well as researchers and directors of testing. Related important topics intentionally not covered by this text include various aspects of implementing placement and exemption programs, academic advising, and providing students with necessary information. Efforts were made in preparing this text to avoid the technical approach wherever possible.

This text consists of seven chapters. The first chapter deals with the statement and nature of the problem, the second provides a somewhat technical rationale for the models to be discussed, and chapters 3 through 6 present, respectively, four classes of alternate treatments, assignment, placement, selection and exemption, and twelve derived models under these classes. Chapter 7 presents the conclusions and implications. An annotated bibliography containing approximately 80 sources is a valuable asset of this text.

Overall, the author of this text has done an admirable job in attempting to take widely diverse educational practices and programs and structuring them into a rational framework that could aid practitioners. There is no doubt that this type of endeavor has been long needed, especially in view of the increased access to higher education over approximately the last decade. As discussed in Chapter 1, this almost unlimited free access has greatly increased the heterogeneity of students attending higher education institutions, and for various reasons institutions have developed a keen desire to respond to the individual needs and interests of students. Therefore, accommodating education to individual differences has received and will continue to receive greater emphasis. A second stated purpose of this text was to describe the educational rationale and technical characteristics of placement, exemption, and closely related models and to help clarify the relationship among them. This is not only a most ambitious purpose, but almost an impossible one to attain considering the complexity of measurement and evaluation and the variance between measurement theory and practice. In Chapter 2 the author states that "decision theory does suggest a general framework for considering problems of alternative educational treatments and it does focus attention on the technical questions that have to be dealt with." Yet, as noted by the author, there are severe restrictions to decision theory. In discussing one of the models of "assignment" in Chapter 3, i.e., "method variation," the purpose of which is to consider ways of identifying the trait-treatment interactions that can make it possible to adapt instruction systematically to individual differences, the author mentions the limitations of current research and states that this is still a research strategy and not an educational strategy. It is this type of problem that diminishes to some extent the usefulness of some of the models presented in this text, for example, in matching students with teachers

with certain characteristics (model 2). Despite this important limitation, as well as lack of adequate consideration of relevant significant psychometric problems, the models presented in Chapters 3 through 6 represent an outstanding attempt to bring some rational structure out of extreme diversity. Chapter 7, the conclusions, is an excellent discussion of the implications of these models in view of our educational setting and some of the current problems such as articulation.

This text should be of immense value, not only to developing institutions, but to well-established institutions with placement and exemption programs. Its primary value lies not in how to develop and implement such programs, which is very much needed but beyond the scope of this book, but in giving institutions a framework upon which to proceed in developing such programs or a framework upon which to evaluate their existing programs. The author has, indeed, accomplished another purpose by reviewing extensively the relevant research literature, and even though much of the research presented cannot be considered as significant research, it nevertheless can be of help to individuals in getting a feel for the "lay of the land." Fragmentation of placement and exemption programs within institutions is a common symptom; this text can be profitable in the improvement of such programs with the end result of better meeting the needs of students.

HENRY MOUGHAMIAN  
*City Colleges of Chicago*









1018  
8/9



EDUCATIONAL and  
PSYCHOLOGICAL

# MEASUREMENT

- W. SCOTT GEHMAN, *Editor*  
GERALDINE R. THOMAS, *Managing Editor*  
WILLIAM B. MICHAEL, *Editor, Validity Studies and Computer Programs*  
JOAN J. MICHAEL, *Assistant Editor, Validity Studies and Computer Programs*  
MAX D. ENGELHART, *Book Review Editor*  
LEWIS R. AIKEN, JR., *Assistant Book Review Editor*  
FREDERIC KUDER, *Editor Emeritus*

## BOARD OF COOPERATING EDITORS

- |  |  |
|--|--|
| DOROTHY C. ADKINS, <i>University of Hawaii</i>                             | LOUIS L. MCQUITTY, <i>University of Miami, Coral Gables</i>          |
| LEWIS R. AIKEN, JR., <i>University of Illinois</i>                         | HOWARD G. MILLER, <i>North Carolina State University at Raleigh</i>  |
| WILLIAM P. BECHTOLDT, <i>The University of Iowa</i>                        | ROBERT L. MORGAN, <i>North Carolina State University at Raleigh</i>  |
| WILLIAM V. CLEMANS, <i>American Institutes for Research</i>                | HENRY MOUGHAMIAN, <i>City Colleges of Chicago</i>                    |
| DAVID D. COHEN, <i>University of Florida</i>                               | DAVID NOVAK, <i>The Neuse Clinic, New Bern, N. C.</i>                |
| ANTHONY J. CONGER, <i>Duke University</i>                                  | ELLIS B. PAGE, <i>University of Connecticut</i>                      |
| LEWIS A. DAVIS, <i>Research Triangle Institute</i>                         | NAMBURY S. RAJU, <i>Science Research Associates, Inc.</i>            |
| WILLIAM A. EDGERTON, <i>Performance Research, Inc.</i>                     | BEN H. ROMINE, JR., <i>University of North Carolina at Charlotte</i> |
| WILLIAM Y. GLASS, <i>University of Colorado</i>                            | THELMA G. THURSTONE, <i>University of Montana</i>                    |
| WILLIAM P. GUILFORD, <i>University of Southern California, Los Angeles</i> | WILLARD G. WARRINGTON, <i>Michigan State University</i>              |
| WILLIAM A. HORNADAY, <i>Babson College</i>                                 | JOHN L. WASIK, <i>North Carolina State University at Raleigh</i>     |
| WILLIAM E. HORROCKS, <i>The Ohio State University</i>                      | KINNARD WHITE, <i>University of North Carolina at Chapel Hill</i>    |
| WILLIAM J. HOYT, <i>University of Minnesota</i>                            | JOHN E. WILLIAMS, <i>Wake Forest University</i>                      |
| WILLIAM D. JACOBSON, <i>University of Virginia</i>                         | E. G. WILLIAMSON, <i>University of Minnesota</i>                     |
| WILLIAM C. JOHNSON, II, <i>Jackson State University</i>                    |  |
| WILLIAM G. KATZENMEYER, <i>Duke University</i>                             |  |
| ROBERT E. LANA, <i>Temple University</i>                                   |  |
| FREDERIC M. LORD, <i>Educational Testing Service</i>                       |  |
| LEWIS LUBIN, <i>Navy Medical Neuropsychiatric Research Unit, San Diego</i> |  |

VOLUME THIRTY-FIVE, NUMBER FOUR, WINTER 1975



## GROUP SIZE EFFECTS IN EMPLOYMENT TESTING<sup>1</sup>

JOSEPH M. HILLERY AND STEPHEN S. FUGITA

University of Akron

Effects of the number of individuals (1 to 10) coacting while taking two standardized motor performance tests were examined. Scores on the manual and finger dexterity sections of the General Aptitude Test Battery were collected from two state employment agencies for 2,261 actual applicants. Increases in aptitude scores corresponding to increases in group size were predicted based upon the summation hypothesis of social facilitation theory. Results indicated a group size effect with performance appearing to increase somewhat linearly with increases in number of coactors. The implications for social facilitation theory and the interpretation of tests administered in a group setting were discussed.

SINCE the publication of Zajonc's (1965) paper drawing on Hull-Spence Theory (e.g., Spence, 1956) to reconcile the conflicting findings in the social facilitation literature, a considerable amount of research has been conducted. Much of the published research supports Zajonc's hypothesis (cf. Zajonc, 1972). Zajonc proposed that the presence of others increases general arousal or generalized drive, and hence, according to the multiplicative drive law, enhances the probability that dominant responses will be emitted. If the dominant responses are correct, as in simple, well-learned or instinctual behaviors, performance will be improved. If, on the other hand, the dominant responses are incorrect as is likely if the behavior is poorly learned or complex, performance will be impaired since the probability of incorrect responses being emitted will be greater.

An implication of an extended version of this theory is that as the

<sup>1</sup> The authors would like to acknowledge the generous assistance of Geoffrey Johnson of the Michigan Employment Security Commission and Frank Lewandoski and Elwood Ziegler of the Ohio Bureau of Employment Services for their help in data collection. Also helpful were Richard H. Haude's, Paul C. Rosenblatt's, and Kenneth N. Wexley's comments on an earlier draft.

number of others present increases, arousal may concomitantly increase and hence, the probability that dominant responses will be emitted may be similarly increased. This summation hypothesis, in the audience paradigm, has been discussed by Weiss and Miller (1971). Although summation is a seemingly simple way to manipulate drive, few social facilitation studies have focused on the group size factor (Dorrance and Landers, 1974). Brenner (1974a) using the Psychological Stress Evaluator (an instrument which monitors voice patterns believed to be associated with arousal in the central nervous system) reports that arousal increases as a power function of audience size in a public speaking situation. His audience group sizes were 0, 2, 8, and 22 spectators. Dorrance and Lander's experiment also partially supported the hypothesis that increases in activation are positively related to increases in audience size. Furthermore, they report an audience size  $\times$  task type interaction using two different kinds of motor performance tasks.

In an early investigation of the effects of an audience on motor and cognitive performance, Gates (1924) reported the possibility of a slight "stimulating" effect with a large audience (27 to 37 members) on a word naming task. More recently, Burwitz and Newell (1972) report that on a novel motor skill where subjects coacted alone, in dyads, or in tetrads, there was no difference between the alone and dyad conditions, but both were significantly superior to tetrads. Martens and Landers (1969) examined performance on a simple, well-learned muscular endurance task under one of three group size conditions; alone, in pairs, or in groups of four. Results showed that coacting individuals in tetrads performed significantly better than those alone or in dyads. Martens and Landers (1972) also ran an experiment which examined performance during the acquisition of a complex motor skill. Group size was again manipulated, this time with four conditions; alone, dyads, triads, and tetrads. Results generally supported the hypothesis that increasing the number of coactors results in increasing impairment of complex motor performance.

The present study was designed to systematically test the group size effect. Ten group sizes which contained from one to ten coacting individuals were examined in a highly standardized employment testing situation. This field situation is a presumably ego-involving one which has serious consequences for those whose performance may be affected by any social facilitation effects. Utilizing the employment test setting also has the advantages of minimizing potential reactive or experimenter effects which might be associated with laboratory research (e.g., Brenner, 1974b) and of providing additional information about

the robustness and generalizability of the social facilitation phenomenon. The experimental tasks were simple, standardized motor tests.

### *Method*

#### *Task*

Scores on the manual and finger dexterity sections of the General Aptitude Test Battery (GATB) developed by the United States Employment Service (Dvorak, 1947) and used extensively by state employment offices throughout the country were collected. Four speed tests which measure the two dexterity aptitudes were used. These were the Place and Turn tests which index manual dexterity and the Assemble and Disassemble tests which index finger dexterity. Both the Place and Turn tests utilize a pegboard divided into two sections. In the Place test, the examinee removes two pegs simultaneously, one in each hand, from holes in the upper part of the board and inserts them into the corresponding holes in the lower part of the board. In the Turn test, the examinee removes a peg from a hole, turns the peg over, and returns it to the hole from which it was taken.

Both Assemble and Disassemble tests use a small rectangular board containing 50 holes and a supply of small metal rivets and washers. In the Assemble test, the examinee takes a rivet from a hole with one hand while removing a small washer from a vertical rod with the other hand. He then puts a washer on the rivet and inserts the assembled piece into the corresponding hole in the lower part of the board. In the Disassemble task, the examinee removes the metal rivet and washer in the reverse manner from the Assemble operation.

#### *Subjects and Procedure*

Scores on the manual and finger dexterity section of the GATB were collected from job applicants to two state employment agencies, Michigan ( $N = 1783$ ) and Ohio ( $N = 428$ ). Arrangements were made with the testing sections of both state agencies to collect data in a specified manner. Branch offices in urban areas of Michigan and the Akron-Youngstown region of Ohio cooperated in the data collection. Experienced test administrators who had been trained by their respective states to give the GATB in a highly standardized manner administered the instrument. Assignment of individuals to group sizes was essentially random in as much as each branch office had prescribed times to administer the test. Thus, the instrument was given to that group of individuals who arrived after the last administration but



before the administration in question. In the Michigan sample, each test administrator was asked to send 75 test record cards or those given in a three month time period. In the Ohio sample, the procedure was modified somewhat to insure that an adequate number of persons tested alone was included in the total sample. Branch offices were asked to alternately send the scores of persons tested individually and for the next group of persons tested regardless of the number of people in the group. Data were also gathered to determine whether any potential group size effect might interact with the sex, age, educational level or race of the applicant. The GATB is administered with the applicants seated around a large rectangular table. Thus, they can see and hear each other as they work on their individual tests.

### Results

Since the data from the two samples should have been and were similar, they were combined for purposes of data analysis. The number of applicants in each group size condition, cell means and standard deviations are presented in Table 1. The ANOVA indicated a clear group size effect with an upward trend in aptitude scores which corresponds with increases in number of coactors in both manual ( $F = 4.54$ ,  $df = 9/2251$ ,  $p < .001$ ) and finger dexterity ( $F = 2.88$ ,  $df = 9/2251$ ,  $p < .002$ ). Tests for linearity indicated that some predictability is afforded by the linear rule for both manual and finger dexterity ( $F = 36.21$ ,  $df = 1/2251$ ,  $p < .001$ , and  $F = 17.04$ ,  $df = 1/2251$ ,  $p < .001$  respectively). The tests for deviation from linearity were nonsignificant for both dependent variables. Sex, age, and educational level of the applicant did not interact with group size on either the manual or finger

TABLE 1  
*Aptitude Test Scores as Related to Group Size*

Group Size	N	Manual Dexterity		Finger Dexterity	
		$\bar{X}$	SD	$\bar{X}$	SD
1	146	96.88	23.38	95.48	21.51
2	125	94.02	24.37	95.72	21.21
3	219	99.33	23.30	95.93	21.77
4	292	98.37	23.44	96.08	20.25
5	290	100.17	23.47	94.97	22.15
6	320	101.34	25.14	96.31	21.23
7	277	102.92	21.96	98.36	21.38
8	212	103.63	22.88	98.83	21.08
9	120	108.36	21.79	103.27	21.14
10	260	103.39	20.96	100.76	20.74
Whole Sample	2261	100.94	23.31	97.35	21.32

dexterity variables. Race of applicant did interact with group size on manual dexterity ( $F = 2.14$ ,  $df = 9/2231$ ,  $p < .05$ ). Inspection of the relevant means indicated that whites exhibited more of the summation social facilitation effect. This interaction was not significant for finger dexterity.

### *Discussion*

These data from an ongoing field setting support predictions consistent with social facilitation theory about the effects of coaction. On simple manual and finger dexterity tasks, subjects performed at a higher level when they coacted in larger groups. The relationship between group size and performance appears to be somewhat linear with the tested group sizes (however, there does seem to be a downturn in groups of ten). On first inspection, this appears to be contrary to Brenner's finding that with group sizes of 0, 2, 8, and 22 spectators, subjects' arousal increased as a power function. However, there are several important differences between his study and the present one. This study's range of group sizes was more limited albeit there was complete sampling within that range. In addition, Brenner's social facilitation situation was an audience type which focused on the mediating mechanism of arousal as opposed to the present coaction setting which examined performance. Since the race  $\times$  group size interaction was only significant on one dependent variable, no conceptual explanation is proposed.

Furthermore, these data are important not only because of their relevance to social facilitation theory, but also because they provide empirical support for Guion's (1965) contention that group as opposed to individual testing may appreciably change the character of a test. The mean differences across disparate group sizes may be of sufficient magnitude to warrant attention in interpreting scores. This might be particularly important with GATB-type instruments in view of the fact that the norms are established with the multiple cutoff method. An individual is considered qualified for a particular job only if he meets the minimum score on each of the key aptitudes (Manual for General Aptitude Test Battery, 1967).

### REFERENCES

- Brenner, M. Stagefright and Steven's Law. Paper presented at the Annual Meeting of the Eastern Psychological Association, Philadelphia, 1974. (a)
- Brenner, M. Mere presence is a Hullian plot. Unpublished manuscript, Ohio State University, 1974. (b)

- Burwitz, L. and Newell, K. M. The effects of the mere presence of coactors on learning a motor skill. *Journal of Motor Behavior*, 1972, 4, 99-102.
- Dorrance, P. D. and Landers, D. M. Social facilitation and drive summation as determined by audience size and evaluative task dimensions. Unpublished manuscript, University of Washington, 1974.
- Dvorak, B. J. The new USES General Aptitude Test Battery. *Journal of Applied Psychology*, 1947, 31, 372-376.
- Gates, G. S. The effect of an audience upon performance. *Journal of Abnormal and Social Psychology*, 1924, 18, 334-342.
- Guion, R. M. *Personnel testing*. New York: McGraw-Hill, 1965.
- Martens, R. and Landers, D. M. Coaction effects on a muscular endurance task. *The Research Quarterly*, 1969, 40, 733-737.
- Martens, R. and Landers, D. M. Evaluation potential as a determinant of coaction effects. *Journal of Experimental Social Psychology*, 1972, 8, 347-359.
- Spence, K. W. *Behavior theory and conditioning*. New Haven, Conn.: Yale University Press, 1956.
- U.S. Department of Labor. *Manual for the General Aptitude Test Battery*. Washington: U.S. Government Printing Office, 1967.
- Weiss, R. F. and Miller, F. G. The drive theory of social facilitation. *Psychological Review*, 1971, 78, 44-57.
- Zajonc, R. B. Social facilitation. *Science*, 1965, 269-274.
- Zajonc, R. B. Compresence. Paper presented at the Annual Meeting of the Midwestern Psychological Association, Cleveland, 1972.

## HIGHEST ENTRY HIERARCHICAL CLUSTERING

LOUIS L. McQUITTY and VALERIE L. KOCH

University of Miami

This paper develops and illustrates the most rapid method yet reported for hierarchically clustering the  $n$  objects of a matrix which portrays the interrelation of every object to every other object, where  $n$  equals any number up to 1000 and even larger. Results compare favorably with those from other excellent methods.

CONCENTRATED Hierarchical Clustering is a rapid and simple method of clustering large numbers of objects (1,000 or more) based on indices of interassociations between the objects (McQuitty and Koch, 1975). The method of this paper was designed to be even simpler and faster.

### *Assumptions in Relation to Time and Effort in Analysis*

Both methods use the concept of a *reciprocal pair* of objects. Two objects,  $i$  and  $j$ , in a matrix of interassociations between objects, are reciprocal if, and only if, Object  $i$  is highest in Column  $j$  and Object  $j$  is highest in Column  $i$ .

Concentrated hierarchical clustering assumes that a reciprocal pair of objects is indicative of a cluster. The current method applies this assumption to the fact that the highest entry in a matrix is reciprocal; if Pair  $i j$  is highest in a matrix, Object  $i$  is highest in Column  $j$  and Object  $j$  is highest in Column  $i$ . Consequently, the current method assumes that the highest entry in a matrix is indicative of a cluster.

An operational difference in the two methods pertains to the fact that a matrix can have more than one reciprocal pair. The earlier method examines the highest entry within every column to determine if it is reciprocal. This takes time and effort. The current method

makes a classification (or classifications in the case of a tie) on the basis of the highest entry only. It reduces the matrix by one object for each pair of objects classified and repeats the operation in terms of the highest entry in the reduced matrix.

The current method saves time and effort with respect to another assumption. When the two objects of a reciprocal pair are classified together, only one of them is removed from the matrix (in order to reduce the size of the matrix), and the other one is retained in the matrix in order that other objects can join those already classified.

The former method assumes that Object  $i$  of the reciprocal pair  $ij$  is the better representative of the initial cluster for retention in the matrix if and only if Object  $i$  exceeds Object  $j$  in being most like other objects remaining in the matrix; it is most like more other objects than is Object  $j$ .

To determine which of two objects,  $i$  or  $j$ , meets the above criterion takes time and effort. With an emphasis on huge matrices and speed of analysis, the current method assumes that either member of the highest entry is an adequate representative of the reciprocal pair for the purpose of classifying other objects with it. One of them is chosen by chance.

The above two alternative assumptions (one for the current method and the other for the former method), about which member to retain in the matrix, were investigated in the earlier study. The results substantiated the adequacy of the assumption which is being applied in the current method (McQuitty and Koch, 1975).

### *Definitions Which Generate the Methods*

The earlier method was developed out of a relatively permissive definition of types. A type is a category of objects of such a nature that every object in the category is reciprocal with one or more other objects in the category. Object  $i$  is reciprocal with Object  $j$  if Object  $i$  is most like Object  $j$  and Object  $j$  is in turn most like Object  $i$ .

The current method will be generated out of a still more permissive definition of types. A type is a category of objects of such a nature that every object in the category is most like one or more other objects in the category; i.e., in terms of the objects remaining in the matrix at the time the object is classified.

### *Restrictive versus Permissive Definitions*

Whether exacting or permissive definitions should be applied in developing methods for the analysis of data depends on the purpose of the analysis. If a primary purpose is to discard irrelevant data, then a



restrictive definition is desirable. Excellent examples are the definitions by Thurstone (1927a, 1927b, 1927c, 1927d, 1928, 1929, and 1932) in the development of linear scales. However, restrictive definitions assume detailed knowledge as to the manner in which data are in fact interrelated (or need to be interrelated for some justifiable purpose).

If the purpose is to discover through analysis the fashion in which a set of data is in fact interrelated, then initial methods of analyses usually need to be developed out of relatively permissive definitions; otherwise the definitions might be restrictive in such a fashion as to preclude discovering the nature of the interrelationships in the data. The definitions can be made more and more restrictive as applications of the permissive definitions yield insights into the actual concatenations in the data.

If application of a restrictive definition yields in some way questionable results, reversion to a more permissive definition might be helpful. The investigator could then move gradually to more and more restrictive definitions as greater and greater insights into the nature of the interrelationships in the data are realized.

From a societal point of view there is a more fundamental issue: the assumption of a particular kind of relationship may foster its development. The assumption of linear relationships in the assessment of educational achievement, for example, may cause individual differences in intelligence to be more nearly linear than it otherwise would be.

### *The Method*

#### *General*

The highest entry in every column of a matrix is identified, and from amongst them the highest entry in the matrix is identified. The two objects between which the highest entry mediates are classified together. The row and column of one of them is selected by chance and removed from the matrix. The highest entry in the reduced matrix is identified, the objects between which it mediates are classified together. One of them is chosen by chance and its column and row are removed from the matrix. The process is repeated until all objects have been classified. Every time an object,  $i$ , is classified with an object,  $j$ , each of them takes with it into the new classification all of the objects with which each has already been classified (if any).

#### *An Illustration*

The method is illustrated with a matrix which is difficult to analyze by some methods because it yields many ties by some methods (Mc-

Quitty, Price, and Clark, 1968). This matrix was used in illustrating Concentrated Hierarchical Analysis, mentioned above.

The matrix is shown in Table 1. The highest entry in every column is underlined. If there is a tie for the highest entry in a column, all tied entries are underlined, as in Column C, where the highest entry is 34.

The highest entry in the matrix is isolated. It is 34, and appears in Columns C, F, and P. The first column and the first row in which the highest entry occurs are designated  $i$  and retained, and the other row(s) and column(s) in which it occurs are designated  $j$  and removed from further analysis. Chance retention and removal, in this approach, is achieved by using chance to assign the objects to their positions in the matrix. The first row and column, C, in which 34 occurs is designated  $i_1$  and the next two rows and columns, F and P, in which it occurs are each designated  $j_1$ . Using the new designation,  $i_1$  is most like  $j_1$ , and  $j_1$  is most like  $i_1$ , irrespective of which  $j_1$  is intended. The columns and rows for both  $j_1$ 's were removed by crossing them out in the matrix. The lines which mark out Rows F and P were terminated in the column of Step 1 to show that they were removed in Step 1. This action completed Step 1, as shown in the table.

Step 2 is a repetition of Step 1, just outlined, except that it applies to the reduced matrix. The highest entry in each column was underlined. There were changes only to the extent that Rows  $j_1$  (F and P), which were removed, thereby changed the highest entries in some retained columns; the removal of Row F, for example, removed some of the entries in Columns, C, I, K, and S, and the new highest entries had to be identified and underlined (except for Column I which had three highest entries and only two of them were removed).

The highest entry in the reduced matrix was identified. It is 33 in Row C—Column I and Row I—Column C. Row and Column I were designated  $j_2$  and removed by marking them out in the matrix. Row and Column C were designated  $i_2$  and retained in the matrix.

No further operations were performed on Table 2. Otherwise to follow the above description would have been difficult. The complete analysis is shown in Table 2.

The removal of Row I, removed the highest entries in Columns C and M. The highest entries in the reduced matrix for these columns were underlined. They are 32 in Row Q—Column C and 28 in Row A—Column M.

In Step 3, the highest entry in the reduced matrix is 32 in Row C—Column Q and Row Q—Column C. Row and Column Q were designated  $j_3$  and removed by marking them out, and Row and Column C were designated  $i_3$  and retained.

Step 4 contains a new kind of tie. The highest entry in the reduced

TABLE 1\*  
Illustrating the First Two Steps of Highest Entry Hierarchical Clustering

		2		I				$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$		$j_1$		$j_1$							
								$i_2$		$i_1$		$j_2$											

\* Data from McQuitty, Price, and Clark, 1967.



matrix is 31, it mediates between Objects (1) C and N, (2) J and T, and (3) K and S. In the previous tie, one object, C, was involved in all of the tied values. Objects C, J, and K were designated  $i_4$ ,  $i_4'$ , and  $i_4''$ , respectively, and retained. Objects N, T, and S were designated  $j_4$ ,  $j_4'$ , and  $j_4''$ , respectively, and removed by marking them out in their rows and columns.

Step 5 shows 30 to be the highest entry in the reduced matrix. It mediates between C and K on the one hand and between G and J on the other, with C and G thus designated  $i_5$  and  $i_5'$ , and retained, and K and J designated  $j_5$  and  $j_5'$  and removed.

Step 6 yields A and C highest with an entry of 29; A was designated  $i_6$  and retained. C was designated  $j_6$  and removed.

Step 7 yields A and M highest with an entry of 28; A was designated  $i_7$  and retained, and M was designated  $j_7$  and removed.

Step 8 yields B and O and D and L with the highest entry of 27. B and D were designated  $i_8$  and  $i_8'$ , respectively, and retained. O and L were designated  $j_8$  and  $j_8'$ , respectively, and removed.

Step 9 yields B and G with the highest entry of 26. B was designated  $i_9$  and retained, and G was designated  $j_9$  and removed.

Step 10 yields the highest entry of 25 for AE on the one hand and BD on the other. A and B were designated  $i_{10}$  and  $i_{10}'$ , respectively, and retained. E and D were designated  $j_{10}$  and  $j_{10}'$ , respectively, and removed.

Step 11 yields the highest entry of 20 for A with B, B with H, and B with R. A was designated  $i_{11}$ . B was first designated  $j_{11}$  and then  $i_{11}'$ . H and R were each designated  $j_{11}'$ .

The above step introduces two novel situations: (1) one object, B, is classified as both a  $j$  and an  $i$ ; and (2) the step is terminal. Whenever an object is classified both  $i$  and  $j$ , it is removed; the  $j$  classification dominates. Objects B, H, and R were removed and Object A was the only one retained, and it had already been classified. The classification terminates when all objects have been classified.

### *The Hierarchical Structure*

The classification structure is shown in Figure 1. It can be constructed from either Table 2 or the description in the steps of the analysis as outlined above. Step 1, for example, classified C with each F and P under a score of 34. An asterisk was placed under each C and F in Step 1 of Figure 1 to show that C joined F in this step, and an asterisk with a prime was placed under each C and P to show that they joined one another in this step. The prime shows that a tie occurred; the two objects with the asterisk and prime tied with the two objects





with the asterisk only. In case of three pairs in a tie, the members of the third pair are designated by an asterisk with two primes as shown in Step 4.

Step 2 classified I and C together under a score of 33, and C, in accordance with the prescribed procedure, took with it into the hierarchical structure the objects (F and P) with which it had already been classified.

In the course of an analysis, a classification sometimes occurs between two objects which have not previously been classified. In cases like this their initial portrayal is on a separate sheet from those objects already classified. They are later joined to another structure if, and only if, one of their objects classifies with an object of another structure.

Building the classification structure proceeded in the above fashion until all objects were classified from the right hand side of the structure to the left hand side according to the size of the scores which joined them to the initial structure (except for ties); the order is arbitrary for tied scores.

### *Proof of the Method*

Let any three objects,  $i$ ,  $j$ , and  $k$ , be so chosen that they are directly associated in a classification resulting from the above kind of analysis:  $i$  is highest with  $j$ ,  $j$  is highest with  $i$ ;  $j$  is also highest with  $k$ , and  $k$  is highest with  $j$ . By definition,  $i$  and  $j$  belong to the same type, and likewise  $j$  and  $k$  belong to the same type. Therefore,  $i$ ,  $j$ , and  $k$  belong to the same type, and all objects associated in a cluster belong to the same type.

An exception to the above conclusion could derive from a classification based on only two objects in a matrix. Consequently, this kind of a classification has no inherent validity. An exception could not derive in any other way. However, the validity of classifications generally becomes less as the size of the matrices on which they are based decreases and as the size of indices of association decrease.

### *Expediting the Method*

The method can be expedited by removing irrelevant data. Irrelevant data are those entries which are not used in classifying objects.

Which data are irrelevant can be estimated, initially. After the highest entry in every column has been isolated, the lowest of these is determined and labeled L. The L entry of the current data is 23 and occurs in Column H, as shown in Table 1. A tentative estimate is made

that all entries lower than L are irrelevant. They are removed from the analysis for the time being.

The reduced set of data is then analyzed in the fashion outlined above for the complete matrix. The analysis continues as long as the method yields classifications. However, the method may discontinue classifications prior to all objects having been classified.

The discontinuance of classification for the reduced data of the current study would have occurred at the end of Step 10. This is because the classification of the next step for the complete matrix was based on a score of 20 (which mediates between A and B, B and H, and B and R) and the lowest entry retained under the above criterion was 23.

Objects A, B, H, and R were the only objects left unclassified at the end of Step 10. All of them had earlier had entries of 23 or above. Object A, for example, had had an entry of 30 with Object P, but this entry was removed when Object P was removed from the matrix in Step 1.

In order to complete the analysis, the unclassified objects are reassembled in a matrix of only them (Table 3 for the current data) and the expedited method is applied to them in the same fashion as it was to the original matrix. The L entry for the reduced matrix is 20. It occurs in every column and, of necessity, classifies the remaining objects in the same fashion as the original method when applied to all of the data.

Additional reduced matrices could have been realized in the above fashion if required for completing the analysis.

One could make a more conservative estimate for removing irrelevant scores from the initial matrix and thereby avoid the necessity of more than one reduced set of data. In the above case, if the L entry for removing entries had been reduced by 13 percent, from 23 to 20, all scores of 20 or above would have been retained and the first reduced set of data would have been sufficient for completing the analysis.

### *Enhancing the Capacity of the Method*

*Description.* The size of a matrix which can be analyzed can be enhanced if the original matrix can be partitioned into two or more overlapping, but of course smaller, sub-matrices, which are chosen in such a fashion that when they are analyzed separately and the results combined they give the same classifications which would have been obtained if the original matrix could have been analyzed without partitioning.

Classification by the complete method (applied to the original, complete matrix) utilizes a numerical criterion which is decreased every

TABLE 3  
*Illustrating Analysis after Satisfying the First Criterion*

	A	B	H	R
A		<u>20</u>	13	16
B	<u>20</u>		<u>20</u>	<u>20</u>
H	13	20		15
R	16	<u>20</u>	15	

time a classification is completed (i.e., except for ties). The original criterion is the highest entry in the original matrix. The two objects between which it mediates are classified together. One of them is removed from the matrix and the other is retained. The criterion is reduced to the highest entry in the reduced matrix, and the process is repeated. The entire process is repeated until all objects are classified.

The above facts can be utilized to partition a matrix in such a fashion that when the analysis is applied to them it will produce the same results as applying the analysis to the unpartitioned matrix.

The highest entry of every column of the original matrix is underlined. In the case of two or more entries tying for highest in a column, all of them are underlined. The highest entry for each column is listed at the bottom of the column, thus forming a bottom row of highest column entries. The entries of this row are assigned ranks, with the largest of these entries having a rank of one, the next largest rank of two, etc. In the case of a tie, the highest numerical rank involved in the tied entries is assigned to all of the tied entries. For example, if the highest entry in the row is 86, followed by 84, 80, 80, 80, 80, 80, and 75, the assigned ranks are 1, 2, 7, 7, 7, 7, 7, and 8 respectively. The ranks assigned to the entries in the bottom row are assigned to all corresponding values *which are underlined* in the body of the original matrix, but not to corresponding values which are not underlined. For example, if the highest entry of Column 1 is 86, of Column 2 is 84, of Column 3 are 80 and 80, of Column 4 are 80, 80, and 80, and of Column 5 is 75, these numbers would be underlined, and would be assigned ranks of 1, 2, 7, 7, 7, 7, 7, and 8 respectively, but if Column 1 also had entries of 84, 80, and 75, they would not be assigned ranks because they would not be underlined as highest entries in Column 1.

Assume that  $X$  equals the number of objects in the largest matrix which is practical or desirable for some reason to analyze (such as limitation of available facilities). The above ranks can be used to select submatrices of size  $X$  or slightly smaller from a larger matrix. In those cases in which the value  $X$  occurs as a rank for highest column entries, the criterion for the number of objects to be included in the first submatrix is  $X$ . In those cases in which  $X$  does not occur as a rank for

highest column entries (because of ties) the criterion for the number of objects to be included in the first submatrix is reduced to the next smaller rank, numerically, which does occur. Let the criterion (whether it be  $X$  or the next smaller rank which occurs) be designated  $C$ . All objects having highest column entries with ranks of  $C$  or smaller, numerically, are selected for the first submatrix.

The first submatrix is analyzed by the first stage of the method outlined above for expediting the method. The smallest underlined entry (for being highest in a column) is designated  $L$ , and all entries lower than it are removed from the submatrix. The analysis of the first submatrix is completed when all of the remaining entries have been used in classifying objects or have been removed as other objects were classified and removed from the matrix (along with their entries).

When the analysis has proceeded to the above stage (where all entries of size  $L$  or above have been used or removed), at least one object will remain in the first submatrix. It is one of the two objects which participated in the last classification; whenever a pair of objects is classified, one of them is removed and the other is retained in the submatrix (until all objects of the original matrix are classified).

Other objects may also remain in the first submatrix at the above stage. They are the objects which had all of their entries of  $L$  or above removed from the first submatrix by virtue of the entries being with objects which were classified and removed from the submatrix.

All objects left in the first submatrix at the above stage are transferred back with those other objects which were not selected for the first submatrix. These objects constitute a matrix of unclassified objects. If the number of objects is still too large for the available facilities, a second submatrix is drawn and analyzed as outlined above for the first submatrix. The process of selecting and classifying by submatrices can be continued until the number of remaining objects is within the limits of the facilities available. They are then classified without further reduction to a submatrix.

*Proof.* When the first submatrix discontinues to yield classifications, all of the objects which could have yielded a classification under a criterion of  $L$  were available and were classified in exactly the same fashion as if the entire matrix had been analyzed. When a new criterion is specified for the next submatrix, all of the objects for classification under the new criterion are available and in the same fashion as for the complete matrix approach. These latter facts apply to all subsequent submatrices. The analysis of partitioned matrices produces the same results as the analysis of the original unpartitioned matrix.



*Evaluating the Method*

The method is evaluated by comparing it with Concentrated Hierarchical Classification which has already been compared with two other versions and one other method and found to be very promising (McQuitty and Koch, 1975).

*Speed and Capacity*

The *Highest Entry* method is generally more rapid and can analyze larger matrices than the earlier method. This is because it omits certain steps: (a) it does not search for all reciprocal pairs; it uses, instead, only the highest entry in a matrix, and (b) it does not apply a criterion to determine which member of a classified pair should be retained in the matrix for further analyses, it selects one randomly. Empirical evidence that the current method is faster is given by the fact that it required only one more step than the earlier method for analyzing the common set of data used in this and the earlier study, and the steps are executed much more rapidly in the current method. Furthermore, the current method has been improved by a technique for eliminating irrelevant data and by matrix partitioning to increase further its capacity.

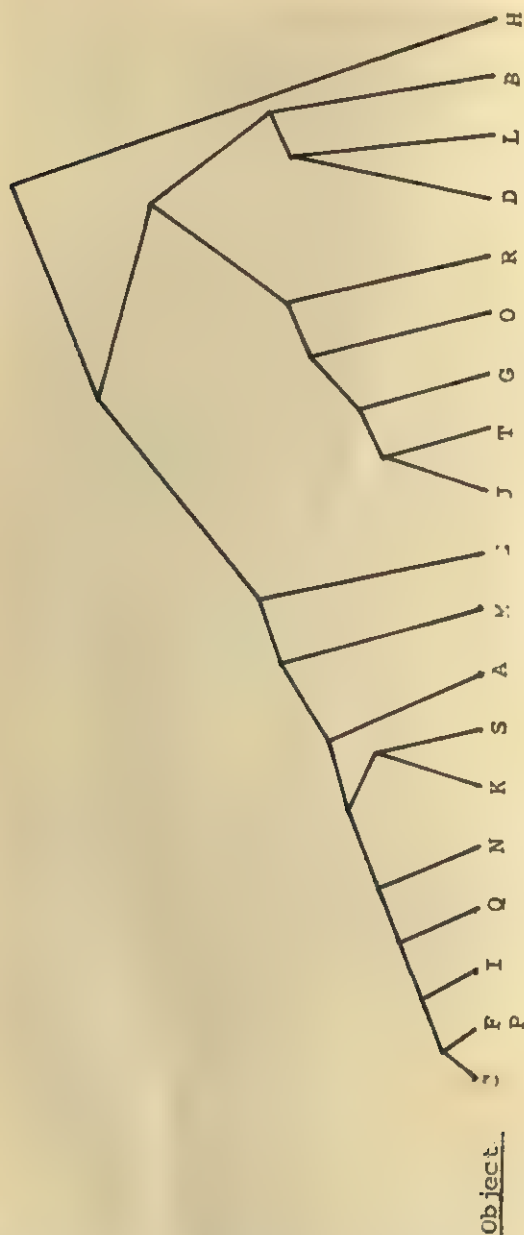
*Reliability and Validity*

A method of classification is generally assumed to be reliable to the extent that its classifications are based on relatively high scores.

Table 4 shows the frequencies and accumulated frequencies of the scores at which classifications were realized in the two methods. Both methods required 19 classifications. The size of the scores at which classifications occurred is identical for the two methods for the first 11 classifications, favors the earlier method slightly for the next six classifications, and favors the current method to a greater extent for the last two classifications. The means of the scores at which the objects

TABLE 4  
*Frequencies and Accumulated Frequencies of Scores at Which Classifications Occurred*

Scores	34	33	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13
Current Method	2	1	1	3	2	1	1	2	1	2	0	0	0	0	3	0	0	0	0	0	0	0
	2	3	4	7	9	10	11	13	14	16	16	16	16	16	19	19	19	19	19	19	19	19
Earlier Method	2	1	1	3	2	1	1	3	2	0	0	0	0	1	0	1	0	0	0	1	0	0
	2	3	4	7	9	10	11	14	16	16	16	16	16	17	17	18	18	18	18	19	19	19



Object.

Scores at which Classifications Occurred

13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34

S 1, 1', 1, 1', 2  
T 2  
E 3  
P 4  
S 5, 6, 7, 8, 9

1" 1"  
2"  
3"  
4"

3'

4  
5

4"  
5"

6  
7  
8

9  
10

10

TABLE 5  
Order of Classification of Objects

Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	<u>16</u>	17	18	<u>19</u>	20	
Earlier Method	<u>1</u>	<u>2</u>	<u>3</u>				<u>1</u>	<u>2</u>				<u>1</u>	<u>2</u>				<u>1</u>	<u>2</u>			
Method	C	F	P	I	Q	N	K	S	A	M	E	J	T	G	O	R	D	L	B	H	
Current Method	<u>1</u>	<u>2</u>	<u>3</u>				<u>1</u>	<u>2</u>				<u>1</u>	<u>2</u>			<u>1</u>	<u>2</u>	<u>1</u>	<u>2</u>	<u>1</u>	<u>2</u>
Method	C	F	P	I	Q	N	K	S	A	M	E	J	T	G	O	B	D	L	R	H	

— " positions which classify differently objects by the two methods.

classify are 28.05 for the current method and 27.94 for the earlier method. These comparisons indicate the methods to be similar with respect to reliability of their classifications.

In studying the validity of the current method, the hierarchical structure from the former study is reproduced in Figure 2 for comparison with that from the current study for the common data of the two studies. The comparison is summarized in Table 5, where the objects are listed from left to right for each method to show the order in which they classified. The order is indeterminable by each method for objects C, F, and P. That they form a cluster of indeterminate order amongst themselves in each method is indicated by the numbers 1, 2, and 3. All other clusters of this kind were for only two objects and are indicated by the numbers 1 and 2 associated with the objects of such clusters. The earlier method produced three clusters of this kind, compared with five for the current method.

In the above cases where the orders were undetermined by the method, the members of the clusters were arranged to minimize the discrepancy between the orders obtained by the two methods. There were differences between the two methods in the position classification of only two objects as shown in Positions 16 and 19 where each B and R in the current method is three positions removed from where they are in the earlier method. The greater speed and capacity of the current method probably compensates in most sets of data for errors (if any) which might result from the current method.

### Conclusions

*Highest Entry Hierarchical Classification* is the most rapid method thus far developed for hierarchical clustering of matrices of inter-associations between objects of all sizes up to  $1,000 \times 1,000$  and even larger, and it compares favorably with the best of methods in terms of reliability and validity.

## REFERENCES

- McQuitty, L. L., Price, L., and Clark, J. A. The problem of ties in a pattern analytic method. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1968, 28, 9-21.
- McQuitty, L. L. and Koch, V. L. A method for hierarchical clustering of a matrix of a thousand by a thousand. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1975, 35, 239-254.
- Thurstone, L. L. Psychophysical analysis. *American Journal of Psychology*, 1927, 38, 368-389. (a)
- Thurstone, L. L. A law of comparative judgment. *Psychological Review*, 1927, 34, 273-286. (b)
- Thurstone, L. L. Equally often noticed differences. *Journal of Educational Psychology*, 1927, 18, 289-293. (c)
- Thurstone, L. L. The method of paired comparison for social values. *Journal of Abnormal and Social Psychology*, 1927, 21, 384-400. (d)
- Thurstone, L. L. The measurement of opinion. *Journal of Abnormal and Social Psychology*, 1928, 22, 415-430.
- Thurstone, L. L. Theory of attitude measurement. *Psychological Review*, 1929, 36, 222-241.
- Thurstone, L. L. Stimulus dispersions in the method of constant stimuli. *Journal of Experimental Psychology*, 1932, 15, 284-297.

## Q FACTOR ANALYSIS: APPLICATIONS TO EDUCATIONAL TESTING AND PROGRAM EVALUATION

F. STEVENS REDBURN  
Youngstown State University

The potential use of *Q* factor analysis for evaluating and designing educational and clinical programs is discussed and illustrated. The advantages of this technique in comparison to normative measurement employing prevalidated or *a priori* scales include additional richness of theoretical insight, discovery of the structure as well as the content of the individual's thinking, relative independence from prior conceptualization, and efficiency in gathering detailed information quickly. *Q* factor analysis is most appropriate for use in clinical or educational situations where available typologies or scales seem inadequate, where the psychological dynamics of learning or treatment are not well understood, or where it is desirable to avoid anticipating the precise direction and character of program impact. Several practical applications of the technique are suggested.

VIRTUALLY all educational measurement and testing focuses on normative comparisons or the movement of subjects relative to prevalidated or *a priori* scales. Although the alternative perspective offered by *Q* methodology, and especially by *Q* factor analysis, is well known (Stephenson, 1953; Kerlinger, 1964), its potential for use in evaluating and designing a great variety of educational and clinical programs has been neglected.

The following brief description of a study of perceptions and attitudes held by a group of college urban interns and changes in their patterns of thinking following the internship will suggest various uses of the *Q* factor analysis methodology. Although the findings of the study have been quite provocative to the faculty associated with this program, the focus here is not on the substantive results but on the potential of *Q* as a testing procedure that, while possessing limitations



of its own, avoids certain pitfalls associated with traditional scaling approaches.

*Rationale for the Use of Q Factor Analysis to  
Measure Educational Program Impacts*

In *Q* methodology, subjects sort (i.e., rank-order) a series of statements according to an abstract criterion (often an agree-disagree dimension).<sup>1</sup> The ordering of statements may then be examined for its meaning relative to a theory or relative to other evidence of the individual's state of mind. Factor analysis adds to the power of this technique by permitting the identification of clusters of individuals who have performed similarly. First the *Q*-sorts of all individuals under study are intercorrelated. The resulting *Q* matrix is then factored to locate clusters of respondents.<sup>2</sup>

The interpretation of the ordering of statements typical of a particular factor is essentially a subjective and imaginative enterprise; it is, therefore, useful on occasion to have two or more individuals independently label and describe the factors. Where there are differences in interpretation, their discussion will often generate additional richness of insight, contributing to the value of the technique as a means of exploration and discovery.

It is the latter purpose which *Q* factor analysis best serves. It is most appropriate for use in clinical or educational situations where available typologies and scales seem inadequate, where the psychological dynamics of learning or treatment are not well understood, or where it is desirable to avoid anticipating the precise direction and character of program impact. In short, this measurement approach is invited in most if not all small group clinical and educational program categories.

Why is *Q* factor analysis more likely to offer fresh insights into the behavior of individuals? In part, this is so because it is a highly efficient technique for quickly gathering detailed information about someone's

<sup>1</sup> In the experiment to be used here for illustration, this is done by sorting 51 statements, each one printed on a separate card, into 11 ranks to produce a prescribed distribution which is the same for each respondent and can be modelled as follows:

Score	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
Frequency	4	4	4	5	5	7	5	5	4	4	4

In this distribution, the numbers above the line represent the scale of discrimination as the sorter uses it, and the numbers below the line represent the numbers of statements to be placed at each increment of the scale.

<sup>2</sup> In *Q* factor analysis, person factors rather than test factors are identified. For the distinction, see Stephenson (1953) and Kerlinger (1964).

thinking in a form that lends itself to quantitative comparisons with the thought of others. Specifically, anyone performing a *Q*-sort of  $n$  statements will make, in effect,  $\frac{1}{2}[n(n-1)]$  comparisons between pairs of statements in little more time than is required to respond to  $n$  independent scale items.

The relative rankings of statements often give insight into the structure as well as the content of an individual's thinking. In the typical testing approach there is a further loss of information (as well as a potential gain in information) when a number of items are added together to form a scale. In an examination of the completed *Q* sort, on the other hand, statements are not treated as scale items (although scales are frequently used to structure a *Q* sample); rather, their relative placements are studied in an attempt to penetrate the logic of the individual. If this examination follows a factor analysis of several sorts, then the ordering examined is actually the weighted average for a similarly sorting group of individuals. In either case, unanticipated item juxtapositions are likely to reveal logical or psycho-logical relationships that would be blotted out by treatment of each subject or set of subjects as a cluster of scale scores. It is this relative independence from a priori categorization that is the primary strength of *Q* when dealing with the impacts of ambiguous, complex symbolic exchanges. Such transactions are the essence of many educational and clinical programs.

It must be observed that independence of a priori categorization is not total in this or any methodological strategy. There are two times in the *Q* research process when predefinition imposes limits on discovery. One is during construction of the *Q* sample, when statements are selected according to an implicit or explicit theory. The use of variance designs in sample construction is not only limiting in the respect suggested but useful in helping to insure ecological representativeness (Brown 1970a and 1970b; Brunswik, 1956). The danger of foreclosing discovery is reduced if one remembers that the use of theory at this stage is intended mainly to maximize coverage of the attitudinal domain of interest; in examination of the completed sort, items should not be treated simply as components of scales but as statements that are invested with various meanings by various individuals. For instance, an item referring to the President that measures one individual's political alienation may be reacted to by a second individual solely in terms of his feelings toward the incumbent. If items have no fixed meanings, then their interpretations will only become apparent, if at all, when viewed in the context of the same individual's responses to other statements or stimuli. Employing a variance design will hopefully provide enough of that context to expose that meaning.

The second point at which preconceptions of the researcher stand in the way of discovery is during interpretation of the completed  $Q$  sort or the ordering of statements typical of a factor. Just as the same statement may generate different meanings for different individuals, so may the same arrangements or patterns of statements suggest different logics to different researchers. If one is reminded that a major value of  $Q$  is its use as a tool of discovery, then the possibility of multiple interpretations has a positive as well as a negative side. If, on the other hand, one is concerned about validity, i.e., is anxious to identify an interpretation which is most consistent with a particular theoretical or professional perspective or with evidence from other types of measurement, there are means available for this. The use of two or more independent judges has been suggested already. A second device, which establishes a matrix for comparisons between subjects and pre-established norms, is to include among the  $Q$  sorts prior to factor analysis sorts that have a previously established theoretical interpretation. These may be sorts performed by individuals who are known to have a particular orientation or "dummy" sorts composed to conform to an abstract type.<sup>3</sup> In the illustrative use of  $Q$  to be discussed in the following section, the  $Q$  sorts of instructors have been included to provide points of reference for interpretation of students' thinking. Of course, there is no device or combination of devices that will remove the subjectivity inherent in both the design and interpretation of educational and psychological research.

The following example suggests one possible use of  $Q$  factor analysis as a testing device in educational or clinical programs.

*Changes in Cognition, Affect, and Evaluation by  
Twenty Urban Interns*

Two classes of urban interns, a total of 20 students, completed a 51-item  $Q$ -sort prior to and at the conclusion of their intern experiences. The twenty-week program under study combines 300 hours of employment with a local public agency, a weekly seminar of three hours, and 90 hours of work on an academic project to be used by the employing agency. It is an intensive, 10 credit-hour exposure that is accorded high prestige by students and occasionally leads to permanent employment of the student by a participating public agency. The primary purpose of the program, as seen by the funding agencies, is to stimulate commitment to careers in urban administration. The course is conducted

<sup>3</sup> The inclusion of individuals for whom an orientation has been previously established is suggested by the work of Fred N. Kerlinger on the use of  $Q$  factor analysis to validate a previously identified structure of social attitudes (Kerlinger, 1972).

jointly by three instructors, each of whom also completed the *Q*-sort during the study period.

The *Q* sample is constructed around 10 dimensions or substantive foci, as shown in Table 1. It is better practice for many purposes to narrow the attitudinal domain of interest so as to provide more detailed information and thereby aid interpretation. The objective here is to cast the net broadly in order to learn whether and to what degree various cognitive, affective, and evaluative orientations are changed by the work of the course. Although a factorial design is not formally employed, an attempt has been made to include under each focus both realistic and stereotyping statements.<sup>4</sup>

For purposes of the factor analysis, the "before" and "after" *Q* sorts are treated as performances of separate individuals. One result of this is that students who show more stability will have a greater influence on the factor structure that emerges. It is also possible to plot, in terms of factor loadings, the movement of each student relative to the factor structure.<sup>5</sup> A student may, for instance, move from a high

TABLE 1  
*Substantive Foci of the Urban Interns Q Sort and Number of  
Statements Devoted to Each*

Focus	Number of Statements <sup>a</sup>
1. Nature of Local Decision-Making Role	9
2. Sense of Political Efficacy	6
3. Attitudes toward Careers in Urban Administration or Politics	4
4. Elitism vs. Populism	6
5. Perceptions of Local Power Structure	8
6. Perceptions of Public Officials' Integrity	4
7. Perceptions of Public Officials' Responsiveness	5
8. Estimated Impact of Local Government Policies	4
9. City's Problems and Probability of Solution	6
10. Attitudes toward Local Area	2

<sup>a</sup> Total exceeds 51 due to double counting of three statements.

<sup>4</sup> A copy of the *Q* sample may be obtained from the author. Typical statements are: (1) Local government is a great force for good in most communities; (2) Local government is only one of many forces acting on the city and probably one of the *least* powerful in shaping the urban environment; (3) I think that through a career in local government I could do a great deal to help my community; and (4) I doubt that I will ever know enough to play a major role in the public affairs of my community.

<sup>5</sup> So far as can be discovered, the few previous applications of *Q* factor analysis to longitudinal analysis with the same set of subjects do not exploit its potential as an instrument for detecting the movement of individuals relative to a given factor structure. See, for instance, John M. Butler's study (1972) which employs repeated *Q* sorts and factor analysis to demonstrate that clients of psychotherapy show significant changes in self-ideal correlations. Butler does not factor together *Q*-sorts completed at different stages of therapy. See also Brenner (1972).



loading on factor one prior to the urban internship to a more ambiguous position in which he displays moderate loadings on factors one and three. The interpretation of this movement depends on examination of the statement orderings typical of the factors and the use of evidence drawn from other sources concerning the student's experiences.

The urban internship is apparently a very different experience for different individuals. Six factors were identified.<sup>6</sup> On each of these anywhere from two to eight students were significantly loaded at program entry and from three to seven at program's end.<sup>7</sup> The factor loadings of students were generally stable, with sixteen students loaded on the same factor before and after. Correlations between before and after performances were all positive, ranging from  $+ .16$  to  $+ .73$  with a mean of  $+ .50$ . On the other hand, some students who remained loaded on one factor showed movement relative to other factors.

A few students produced strikingly different performances on the two sorts. In particular, two students who were initially loaded on the first factor dropped off this factor and two students moved onto this factor.

**FACTOR ONE:** Because space does not permit full discussion of the substantive results, analysis will be confined to the most heavily populated of the six factors. This factor is of interest also because of the changes in student loadings referred to above and because the ordering of statements by this group corresponds closely to a frame of mind consistent with what was identified above as a principal objective of the internship program.

Table 2 lists the four statements most strongly agreed with and the four statements most strongly disagreed with by the typical member of this group, excluding one statement on regionalism that received high scores on all six factors.

This group of interns is excited about the possibilities for constructive change through careers in urban government. They are anxious to be part of this process. They reject the most cynical stereotypes of local officials and are, if anything, overoptimistic in estimating the potential impact of government on the city. They are also democratic rather than elitist in orientation and (based on the rankings of statements not

<sup>6</sup> A varimax orthogonal rotation was employed, with a minimum eigenvalue of 1.8 used as the standard for retention of factors prior to rotation.

<sup>7</sup> Persons designated as significantly loaded are those located above  $\pm .35$ , significant according to the Guilford-Lacey expression for the standard error of a zero correlation. Persons designated as defining for a factor are those who meet the above criterion and have all other factor loadings at least  $.15$  lower than their highest loading. Defining persons are those whose  $Q$ -sorts are included when computing the weighted average sort typical of a factor.



TABLE 2  
*Extreme Statements Typical of Factor One with Z-Scores<sup>a</sup>*

Statement	Z-Score
I think that through a career in local government I could do a great deal to help my community.	+1.56
Local government's impact on the design of the city and the solution of its problems is probably greater than it has ever been thanks to improved planning and a huge influx of new funds	+1.54
I think I would enjoy some kind of career in urban administration.	+1.52
If we are to have good government, we must not let power slip from the hands of the people.	+1.37
There is so much easy money in government today that I am afraid a lot of local officials are on the take.	-1.57
Local government is only one of the many forces acting on the city and probably one of the <i>least</i> powerful in shaping the urban environment.	-1.59
The role of the elected local official consists largely of doing favors for one's friends and relatives.	-1.83
I doubt that I will ever know enough to play a major role in the public affairs of my community.	-1.92

<sup>a</sup> The four statements above the line are those receiving strongest agreement from this group, the four below the line are those receiving strongest disagreement.

shown in Table 2) perceive local government to be run basically on the democratic model.

Figure 1 shows the shifts in student loadings on this factor. The plus sign at the right of the diagram indicates the positive loading of one instructor on this factor.<sup>8</sup> The two students moving onto this factor during the program report having highly positive internship experiences. One of these students received a permanent job with his agency; the second volunteered to continue participating in the seminar portion of the program to help orient incoming interns. The clear downward movement of five interns relative to this factor may indicate somewhat dampened enthusiasm and/or growing realism about the impact individuals and governments can have on the urban environment.

As previously suggested, there is not a great deal of movement relative to the factor structure, particularly with respect to factors two

<sup>8</sup> In this instance, the three instructors show distinctly different orientations. Instructor A is a hybrid of factors 1 and 2; instructor B is on factor 2; and instructor C is marginal to factor 4. This lack of consensus no doubt contributes to the variety of meanings that can be extracted from the internship experience and possibly weakens its effect generally.

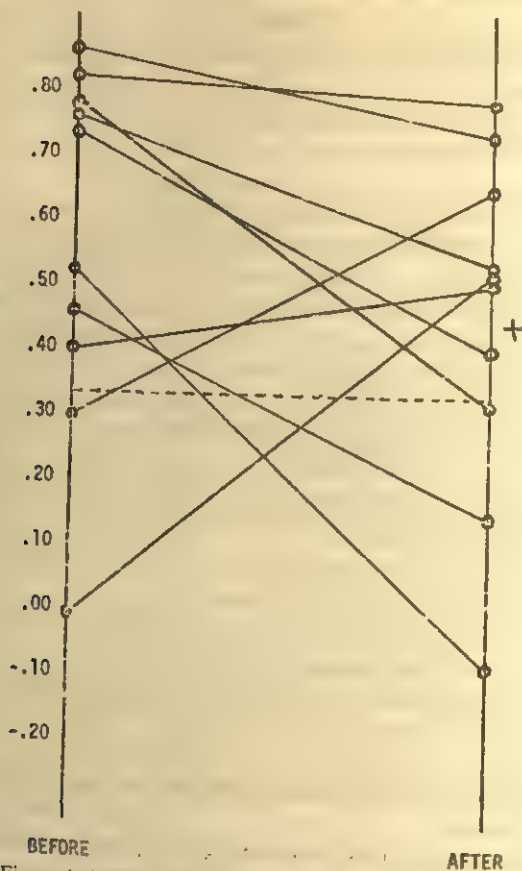


Figure 1. Movement of Students Relative to Factor One.

Shows Before-After comparisons for those loaded before or after at  $\pm .35$  or above. Instructor loading is represented by cross at right.

through six. Perhaps there are clues to the stability of this structure in the apparent focus of each factor. Factor two can be characterized, superficially at least, as a reflection of political liberalism. Factors three and five suggest respectively an outsider (or alienated) and an insider perspective on local affairs. Factor four is more difficult to characterize in an offhand manner. Factor six can be described loosely as the viewpoint of the outwardly mobile, those seeking to leave behind the city and its problems. Such patterns of thinking seem in the main to be rooted in something more than previous academic secondary socialization. If specific cognitions regarding the political and social character of an urban area or the possibilities of altering that

character are, in fact, imbedded in a structure of belief that has been forged during primary socialization and is still being reinforced by significant others, this would explain why even the most intense academic experiences would in many instances have little impact on these specific perceptions (Berger and Luckmann, 1966).<sup>9</sup>

The full elaboration of this line of thought deserves far more space than is available here. This very general review of substantive findings is intended rather to illustrate the suggestive power of *Q* factor analysis as a device for program evaluation. In this case, *Q* analysis obviously raises fundamental doubts about the direction and intensity of the urban internship's short-run impact on cognitions.<sup>10</sup>

### *Possible Applications of Q Factor Analysis to Educational Program Evaluation and Testing*

The following list of possible applications of *Q* factor analysis to educational testing and program evaluation is not meant to be exhaustive and will no doubt suggest other uses. *Q* factor analysis may be used:

1. *To estimate at entry the probability an individual will respond appropriately to a particular treatment or program.* A gradual build-up of experience with various "factors" or "types" would naturally lead to more individualization of curricula or treatments. In this case, *Q* would be a supplement to the judgments of sensitive and experienced professionals. Use of *Q* factor analysis for initial screening would also provide a basis for exclusion of those already possessing the attitude configuration at which it is aimed. (See also use 4.)
2. *For continuous monitoring of short-term movements relative to a factor structure.* This may be either for the purpose of evaluating the impact of program components, e.g., a teaching module or field trip, or to determine the length or type of program needed

<sup>9</sup> Berger and Luckmann suggest that all secondary socialization processes must build on or seek to overcome the subjective reality constructed during primary socialization. To alter an established structure of belief those engaged in secondary socialization must intensify the affective charge of the process through establishment of intimacy and identification or by forging a psychological link between the new reality being presented and the "home" reality produced by primary socialization (Berger and Luckmann, 1966, pp. 130-135; Mead, 1934).

<sup>10</sup> The efficacy of such a program may, of course, be measured in other terms, such as the numbers of students choosing urban careers. It is also possible that major changes in attitude structure have been introduced that will not be manifest until later. Certainly, follow-up testing is called for if the intention is to evaluate the full impact of such a program. The use of control groups will be essential for some testing purposes.

by an individual. The factor structure may be one determined by the current student or client population or it may be defined by a representative group of previous students plus the individual under examination or it may be defined primarily by a group selected for some other purpose. (See uses 5 and 6.)

3. *To specify conflicts of viewpoint among instructors or other program personnel* for the purpose of assessing the impact of these conflicts on students or clients, to resolve such conflicts, or simply to clarify the nature of such conflicts so they can be taken into account or called to the attention of students or clients.
4. *To identify attitude types or factors that are frequently products of primary socialization and so are resistant to change and to identify those which are ordinarily less stable.* (See the discussion of the urban internship program factors in the preceding section.)
5. *To evaluate student performances in order to make normative judgments regarding changes in cognitions.*<sup>11</sup> Norms need not be established prior to testing but instead identified through factor analysis. For instance, if there is a professional consensus on the desirable pattern of thinking then this can be identified by including a set of professionals in the population of sorters. In the absence of such a consensus, of course, no normative judgments are justified.
6. *To identify how the presence of various attitude "types" in a class or treatment group affects the individual's response to the program.* Class peers often play a major role in mediating program effects. By systematically varying the presence and proportions in the class of various attitude "types" it may be possible to learn in a general way the contribution this mediation makes to the program's impact.

### Conclusion

The purpose of this discussion has been to present the strengths of *Q* factor analysis as a technique for use in educational and other

---

<sup>11</sup> Educational testing is often employed as a device for grading. While it is not the purpose of this essay to challenge that use, *Q* factor analysis does raise a conceptual as well as a practical challenge to the relevance of conventional systems of grading student performance in many educational programs. Such systems focus on changes in the substance or quality (e.g., complexity) of students' cognitive orientations. If the symbolic transactions are complex and ambiguous, if the meanings taken from the experience are highly variable and perhaps unique to each individual, then the problems of measuring and modelling the impact of the program are considerable. The appropriateness of attempting to apply common standards of performance to the clients of such programs should be challenged.

programs characterized by complex symbolic transactions. Its general advantage seems to be that it avoids the rigid imposition of a set of categories established prior to testing. The rationale for avoiding such a priori constructions is given by the complexity of both aims and structure that is typical of such programs. Such complexity is itself a reflection of individuals' variety and complexity and the consequent need for programs that will approach one set of objectives for certain individuals and quite another set for others.

The post-industrial future, according to one view, will be characterized by a growing volume of human services in the model of those addressed here, i.e., two-way or multi-sided symbolic interactions where clients become participants who help define the nature of their problems and help work out the solutions (White and Gates, 1974).

If we are to have hope of properly assessing the impacts of such exchanges, then it will be necessary to develop monitoring techniques that avoid stereotyping participants, are sensitive to movements in unanticipated directions, and enrich the possibilities for discovery of a range of participant responses that cannot ordinarily be anticipated in the design of traditional impact measures. *Q* factor analysis is one such technique.

## REFERENCES

- Berger, P. L. and Luckmann, T. *The social construction of reality*. Garden City, N. Y.: Doubleday, 1966.
- Brenner, D. J. Dynamics of public opinion on the Vietnam war. *Science, psychology, and communications: Essays honoring William Stephenson*. New York: Teachers College Press, 1972, 345-380.
- Brown, S. R. On the use of variance designs in *Q* methodology. *The Psychological Record*, 1970, 20, 179-189.
- Brown, S. R. and Ungs, T. D. Representativeness and the study of political behavior: An application of *Q* technique to reactions to the Kent State incident. *Social Science Quarterly*, 1970, 51, 514-526.
- Brunswik, E. *Perception and the representative design of psychological experiments*. Berkeley: University of California Press, 1956.
- Butler, J. M. Self-concept change in psychotherapy. *Science, psychology, and communication: Essays honoring William Stephenson*. New York: Teachers College Press, 1972, 141-171.
- Kerlinger, F. N. *Foundations of behavioral research*. New York: Holt, Rinehart and Winston, 1964.
- Kerlinger, F. N. A *Q* validation of the structure of social attitudes. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1972, 32, 987-995.



- Mead, G. H. *Mind, self, and society*. Chicago: University of Chicago Press, 1934.
- Stephenson, W. *The study of behavior; Q technique and its methodology*. Chicago: University of Chicago Press, 1953.
- White, O., Jr. and Gates, B. L. Statistical theory and equity in the delivery of social services. *Public Administration Review*, 1974, 34, 43-51.

## THE CALCULATION OF RELIABILITY FROM A SPLIT-PLOT FACTORIAL DESIGN

ROBERT L. BRENNAN

State University of New York at Stony Brook

This paper treats the question, "How should one estimate the reliability of school (or classroom) means when persons are nested within schools (or classrooms)?" We begin by reviewing the use of variance components in the estimation of reliability from a randomized block (RB) design. Then we extend this rationale to the estimation of reliability (or generalizability) coefficients in a split-plot factorial (SPF) design with persons nested within schools.

Through the use of variance components from the SPF design, we derive estimates of reliability for schools and for persons within schools. Then we compare the reliability for persons within schools from a SPF design with the reliability for persons from a RB design. Finally, we compare the reliability for schools from a SPF design with the reliability for school means from a RB design.

PERHAPS the most widely used formulas for measuring reliability when a test is administered once to a group of persons are those that can be viewed as providing the average of all possible split-half reliability coefficients for the test. Cronbach (1951) has shown that his Coefficient  $\alpha$ , Kuder and Richardson's (1937) Formula-20, and Hoyt's (1941) reliability coefficient can all be interpreted in this manner, and are, therefore, coefficients of equivalence. In order to calculate any of these coefficients one makes use of sample statistics from a persons by items data matrix; moreover, in calculating Hoyt's coefficient one analyzes the data matrix in the framework of a randomized block factorial analysis of variance design.

In this paper, Hoyt's idea of calculating reliability from a randomized block design is extended to a split-plot factorial design. In its simplest form this design allows for the incorporation of an added dimension into reliability analyses, namely the nesting of persons in

some larger unit, such as schools or classrooms. In the opinion of this author, the experimental model used to collect data for most reliability analyses is usually one in which persons are nested within some dimension; therefore, the split-plot design would appear to be more appropriate than a simple randomized block design. In addition, the split-plot design can be used to provide a basis for estimating the reliability of scores for the units (e.g., schools) within which persons are nested. Such reliability estimates are often needed for statistical analyses in which the unit of analysis is the school or classroom, rather than the person.

In order to introduce the method of calculating reliability from a split-plot design, we begin with a brief description of the use of variance components in the calculation of reliability from a randomized block design. The reader unfamiliar with this technique might consult Lindquist (1953) or Cronbach, Gleser, Nanda, and Rajaratnam (1972) for more detail.

### *Reliability Using a Randomized Block Design*

A randomized block design is a repeated measures design in which the linear model is:

$$X_{ij} = M + P_i + I_j + PI_{ij} + E_{\alpha(i,j)} \quad (1)$$

where

$X_{ij}$  = response of person  $i$  ( $i = 1, 2, \dots, n$ ) to item  $j$  ( $j = 1, 2, \dots, k$ ),

$M$  = population grand mean,

$P_i$  = effect for person  $i$  in the population,

$I_j$  = effect for item  $j$  in the population, and

$PI_{ij}$  = interaction in the population of person  $i$  with item  $j$ , which is confounded with

$E_{\alpha(i,j)}$  = experimental error.

Table 1 provides computational formulas and expected values of the mean squares for this design.<sup>1</sup>

Reliability is, in general, defined as the ratio of the variance of true scores to the variance of observed scores. From (1) it is clear that the variance of the observed scores for item  $j$  is given by:

$$\sigma_{x_j}^2 = \sigma_P^2 + \sigma_E^2 \quad (2)$$

<sup>1</sup> Millman and Glass (1967) and Kirk (1968) provide rules that greatly simplify the determination of expected values of mean squares.

TABLE 1  
Randomized Block ANOVA Table

Source	S.S.	d.f.	M.S.	$E(M.S.)^a$
Between Persons	$(P) - (C)$	$n - 1$	$MS(P)$	$\sigma_E^2 + \sigma_{PI}^2 + k\sigma_P^2$
Within Persons	$(PI) - (P)$	$n(k - 1)$	$MS(w.P)$	
Items	$(I) - (C)$	$k - 1$	$MS(I)$	$\sigma_E^2 + \sigma_{PI}^2 + n\sigma_I^2$
Persons by Items	$(PI) - (P) - (I) + (C)$	$(n - 1)(k - 1)$	$MS(PI)$	$\sigma_E^2 + \sigma_{PI}^2$
Total	$(PI) - (C)$	$nk - 1$		

$$(P) = \sum_i \sum_j X_{ij}^2 \quad (P) = \frac{1}{k} \sum_i (\sum_j X_{ij})^2$$

$$(C) = \frac{1}{nk} (\sum_i \sum_j X_{ij})^2 \quad (I) = \frac{1}{n} \sum_j (\sum_i X_{ij})^2$$

The indicated expected values are for the random effects model. Since experimental error ( $E$ ) is confounded with the persons by item interaction term ( $PI$ ), we will let  $\sigma_E^2 = \sigma_E^2 + \sigma_{PI}^2$ .

where

$\sigma_P^2$  = true score variance (i.e., the population variance for persons), and  
 $\sigma_{E_j}^2$  = error variance for item  $j$ .

Since  $\sigma_{E_j}^2 = \sigma_E^2$  for all  $k$  items, the reliability of a one-item test is given by the intraclass correlation coefficient

$$r_{11} = \frac{\sigma_P^2}{\sigma_P^2 + \sigma_E^2} \quad (3)$$

Now, using Table 1 we find that  $\sigma_P^2$  is estimated by  $[MS(P) - MS(PI)]/k$  and  $\sigma_E^2$  is estimated by  $MS(PI)$ . Thus,

$$r_{11} = \frac{\frac{MS(P) - MS(PI)}{k}}{\frac{MS(P) - MS(PI)}{k} + MS(PI)} \quad (4)$$

$$= \frac{MS(P) - MS(PI)}{MS(P) + (k - 1)MS(PI)} \quad (5)$$

In order to determine the reliability of a test of length  $k'$ , one can use the Spearman-Brown formula in conjunction with (3) or (5), above.

Thus,

$$r_{k'k'} = \frac{k'r_{11}}{1 + (k' - 1)r_{11}} \quad (6)$$

$$= \frac{\sigma_P^2}{\sigma_P^2 + \sigma_E^2/k'} \quad (7)$$

$$= \frac{MS(P) - MS(PI)}{MS(P) + \left(\frac{k - k'}{k'}\right)MS(PI)} \quad (8)$$

One can also determine the reliability of a test of length  $k'$  by making a direct application of the variance components from the randomized block design. From (1) we find that

$$\begin{aligned} V\left[\frac{1}{k'} \sum_i^{k'} X_{i,}\right] &= V(P_i) + V\left[\frac{1}{k'} \sum_i^{k'} E_{i,}\right] \\ &= \sigma_P^2 + \sigma_E^2/k'; \end{aligned} \quad (9)$$

i.e., the variance of the observed score means<sup>2</sup> over  $k'$  items equals the variance of the true scores plus the variance due to errors of measurement. Thus,

$$r_{k'k'} = \frac{\sigma_P^2}{\sigma_P^2 + \sigma_E^2/k'}, \quad (10)$$

which is identical to (7) and, therefore, can be estimated by (8). Note that the error variance in (9) conforms with the usual formula for the variance of a mean. This use of variance components in conjunction with the basic notion of reliability has been extended by Cronbach et al. (1963, 1972) into the Theory of Generalizability. In this theory,  $\sigma_P^2$  in (10) is called the universe score variance, (9) is called the expected value of the observed score variance, and (10) is called the coefficient of generalizability.

When  $k = k'$  in (6), (7), (8), or (10), then we get, what is often called, the reliability of a test of full length:

$$r_{kk} = \frac{MS(P) - MS(PI)}{MS(P)} \quad (11)$$

Formula (11) is best-known as Hoyt's (1941) reliability formula, which is algebraically equivalent to Cronbach's (1951) coefficient  $\alpha$ . When all items are scored in a correct-wrong (1-0) manner, then For-

<sup>2</sup> Here, and elsewhere in this paper, we speak about the reliability of mean scores. The reader should recall that the reliability of mean scores is identical to the reliability of total scores.



mula (11) is also equivalent to Kuder and Richardson's (1937) Formula-20.

### *Reliability Using a Split-Plot Factorial Design*

The above formulas for reliability using a randomized block are well-documented in the literature (see, for example, Stanley, 1971). Here we extend the ideas of the previous section to a split-plot factorial design in which persons are nested within some higher order dimension (which, here, we will call "schools"), persons are crossed with items, and schools are crossed with items. Thus,

$$X_{mij} = M + S_m + P_{i(m)} + I_j + SI_{mj} + IP_{j(i(m))} + E_{o(ijm)} \quad (12)$$

where

$X_{mij}$  = response of person  $i$  ( $i = 1, 2, \dots, n$ ) nested in school  $m$  ( $m = 1, 2, \dots, r$ ) to item  $j$  ( $j = 1, 2, \dots, k$ ),

$M$  = population grand mean,

$S_m$  = effect for school  $m$  in the population,

$P_{i(m)}$  = effect for person  $i$ , nested within school  $m$ , in the population,

$I_j$  = effect for item  $j$  in the population,

$SI_{mj}$  = interaction in the population of school  $m$  and item  $j$ , and

$IP_{j(i(m))}$  = interaction in the population of item  $j$  and person  $i$ , nested within school  $m$ , which is confounded with

$E_{o(ijm)}$  = experimental error.

Note that, in order to simplify the exposition, we assume here that there are an equal number ( $n$ ) of persons nested within each school. Under these circumstances, computational formulas for the analysis of variance in the split-plot design are given in Table 2 with mean squares and expected values of the mean squares given in Table 3.<sup>3</sup>

Using the information in these tables we will express two different reliability coefficients for the same test and compare these coefficients with similar coefficients obtained from a randomized block design. We will not, however, go through the steps involved in deriving the coefficients that emanate from the split-plot design. The derivations proceed in a straight-forward manner by applying the method of variance com-

<sup>3</sup> If the number of persons nested within each school is not the same for all schools then one must obtain adjusted mean squares. If persons were lost for reasons unrelated to the conduct of the experiment, then an unweighted means solution should be used. If the experiment deliberately employed unequal  $n$ 's, then a least squares solution should be used. See Kirk (1968, pp. 204-208, 276-281) for a discussion of these analytic techniques.

TABLE 2  
Computational Formulas for Split-Plot AVOVA

Source	Sums of Squares	Degrees of Freedom
Between Subjects	$(SP) - (C)$	$rn - 1$
Schools	$(S) - (C)$	$r - 1$
Persons within schools	$(SP) - (S)$	$r(n - 1)$
Persons within School $m$	$(SP_m) - (S_m)$	$n - 1$
Within Subjects	$(SPI) - (SP)$	$rn(k - 1)$
Items	$(I) - (C)$	$k - 1$
Schools by Items	$(SI) - (S) - (I) + (C)$	$(r - 1)(k - 1)$
Items by Persons within Schools	$(SPI) - (SP) - (SI) + (S)$	$r(n - 1)(k - 1)$
Items by Persons within School $m$	$(SPI_m) - (SP_m) - (SI_m) + (S_m)$	$(n - 1)(k - 1)$
Total	$(SPI) - (C)$	$rnk - 1$

$$(SPI_m) = \sum_i \sum_j X_{mij}^2 \quad (SPI) = \sum_m (SPI_m) \quad (C) = \frac{1}{rnk} \left( \sum_m \sum_i \sum_j X_{mij}^2 \right)$$

$$(SP_m) = \frac{1}{k} \sum_i \left( \sum_j X_{mij} \right)^2 \quad (SP) = \sum_m (SP_m) \quad (I) = \frac{1}{rn} \sum_j \left( \sum_m \sum_i X_{mij} \right)^2$$

$$(S_m) = \frac{1}{nk} \left( \sum_i \sum_j X_{mij} \right)^2 \quad (S) = \sum_m (S_m)$$

$$(SI_m) = \frac{1}{n} \sum_j \left( \sum_i X_{mij} \right)^2 \quad (SI) = \sum_m (SI_m)$$

TABLE 3  
Mean Squares and Expected Values of Mean Squares for Split-Plot ANOVA

Source	Mean Squares (MS)	$E(MS)^a$
Between Subjects	$MS(b.Subjs.)$	
Schools	$MS(S)$	$\sigma_B^2 + \sigma_{IP}^2 + n\sigma_{SI}^2 + k\sigma_P^2 + nk\sigma_S^2$
Persons within Schools	$MS(Pw.S)$	
Persons within School $m$	$MS(Pw.S_m)$	$\sigma_B^2 + \sigma_{IP}^2 + k\sigma_P^2$
Within Subjects	$MS(w.Subjs.)$	
Items	$MS(I)$	$\sigma_B^2 + \sigma_{IP}^2 + n\sigma_{SI}^2 + nr\sigma_I^2$
Schools by Items	$MS(S \text{ by } I)$	$\sigma_B^2 + \sigma_{IP}^2 + n\sigma_{SI}^2$
Items by Persons within Schools	$MS(I \text{ by } Pw.S)$	$\sigma_B^2 + \sigma_{IP}^2$
Items by Persons within School $m$	$MS(I \text{ by } Pw.S_m)$	

<sup>a</sup> The indicated expected values are for the random effects model. If schools are fixed and items and persons are random, then the expected values remain the same, with the exception of the expected value of  $MS(I)$  which becomes  $\sigma_B^2 + \sigma_{IP}^2 + nr\sigma_I^2$ . Since experimental error is confounded with the items by persons within schools interaction term, we will let  $\sigma_E^2 = \sigma_B^2 + \sigma_{IP}^2$ .

ponents in conjunction with the basic definition of reliability (or generalizability).

### *Reliability for Persons*

The reliability of a test of length  $k'$  for persons within schools is given by:

$$r_{k',k'} = \frac{\sigma_P^2}{\sigma_P^2 + \sigma_E^2/k'}, \quad (13)$$

which is estimated by:

$$r_{k',k'} = \frac{MS(Pw.S) - MS(I \text{ by } Pw.S)}{MS(Pw.S) + \left(\frac{k - k'}{k'}\right) MS(I \text{ by } Pw.S)}. \quad (14)$$

It is interesting to compare (13) and (14) with the reliability that would be obtained if all  $rn$  persons were placed in a randomized block design, thus disregarding the school dimension. Under these circumstances, the resulting reliability, expressed in terms of variance components from the split-plot design, is given by:

$$r_{k',k'} = \frac{\sigma_P^2 + \frac{n(r-1)}{rn-1} \sigma_S^2}{\sigma_P^2 + \frac{n(r-1)}{rn-1} \sigma_S^2 + \frac{1}{k'} \left[ \frac{n(r-1)}{rn-1} \sigma_{SI}^2 + \sigma_E^2 \right]}, \quad (15)$$

which is estimated by:

$$r_{k',k'} = \frac{MS(b.Subjs) - \left[ \frac{rn \ MS(w.Subjs) - MS(I)}{rn - 1} \right]}{MS(b.Subjs) + \frac{k - k'}{k'} \left[ \frac{rn \ MS(w.Subjs) - MS(I)}{rn - 1} \right]}. \quad (16)$$

Algebraic manipulation of (13) and (15) reveals that (13) is less than, equal to, or greater than (15) depending upon whether  $\sigma_P^2/\sigma_E^2$  is less than, equal to, or greater than  $\sigma_S^2/\sigma_{SI}^2$ , respectively. Thus, if one uses a randomized block design to calculate reliability for persons when, in fact, persons are nested within some dimension, such as schools or classrooms, the resulting coefficient will be biased, and, moreover, the direction of bias will be unknown. Since, in most reliability studies, persons are sampled in a stratified random manner, rather than simply randomly, it would appear that Formulas (13) and (14) are preferable to the reliability formulas emanating from a randomized block design.

Using the split-plot design, one can also determine the reliability for persons within any school  $m$ . This is given by:

$$r_{k',k'} = \frac{MS(Pw.S_m) - MS(I \text{ by } Pw.S_m)}{MS(Pw.S_m) + \left(\frac{k - k'}{k'}\right)MS(I \text{ by } Pw.S_m)}, \quad (17)$$

which is algebraically equivalent to the coefficient one would get if the  $n$  persons in school  $m$  were analyzed alone in a randomized block design.

### *Reliability for Schools*

In thinking about the reliability of a test, many of us are so accustomed to thinking about reliability for persons that we overlook the fact that reliability is a generic concept that is applicable to any unit of analysis. Often, in educational and psychological research, the intended unit of analysis is *not* the person. For example, it often occurs in large-scale observational studies that tests are administered to persons within a number of schools, and the unit of analysis for statistical purposes is the school, not the person. Thus, if reliability is to be taken into account in such analyses, one needs the reliability of school means, which can be obtained quite easily from the split-plot design.

Using the method of variance components, one finds that the reliability of a test of length  $k'$ , for schools, is given by:

$$r_{k',k'} = \frac{\sigma_S^2}{\sigma_S^2 + \frac{1}{n}\sigma_P^2 + \frac{1}{k'}\left(\frac{1}{n}\sigma_E^2 + \sigma_{SI}^2\right)}. \quad (18)$$

In general, Formula (18) cannot be expressed very succinctly in terms of mean squares from the split-plot design; however, if  $k' = k$ , then (18) can be expressed as:

$$r_{kk} = \frac{MS(S) - [MS(Pw.S) + MS(S \text{ by } I) - MS(I \text{ by } Pw.S)]}{MS(S)}. \quad (19)$$

It should be noted that, in this case, one must use the method of variance components in order to determine reliability because the Spearman-Brown Formula cannot take into account the fact that  $\sigma_P^2/n$  is part of the variance due to errors of measurement for all values of  $k'$ . This problem is encountered because we are calculating reliability for schools from a design in which the dependent variable is a score for a person nested within a school.

Another way to approach the problem of calculating reliability for schools would be to use a randomized block design in which the dependent variable is an item mean over persons (i.e., item difficulty

level) within a school. In this case, the reliability for schools, expressed in terms of variance components from the split-plot design, is given by:

$$r_{k'k'} = \frac{\sigma_s^2 + \frac{1}{n} \sigma_p^2}{\sigma_s^2 + \frac{1}{n} \sigma_p^2 + \frac{1}{k'} \left( \frac{1}{n} \sigma_s^2 + \sigma_{SI}^2 \right)}, \quad (20)$$

which is estimated by:

$$r_{k'k'} = \frac{MS(S) - MS(S \text{ by } I)}{MS(S) + \frac{k - k'}{k'} MS(S \text{ by } I)}. \quad (21)$$

The only difference between (20) and (18) is in the addition of the term  $\sigma_p^2/n$  in the numerator of (20). Thus, (20) is an upwardly biased estimate of (18). In effect, (20) includes a function of the variance due to persons within schools in what should be the true (or universe) score variance due to schools alone. However, if  $\sigma_p^2$  is close to zero or  $n$  is quite large, the discrepancy between (20) and (18) will be correspondingly small.

### Discussion

The formulas in the previous section were derived under the assumption that all effects are random; however, these formulas are equally valid if persons and items represent random effects and schools represent fixed effects. Under these circumstances, as noted in the footnote to Table 3, the only  $E(MS)$  that is altered is the one for items, and  $MS(I)$  does not affect any of the previous formulas.

Brennan and Kane (1975) provide formulas for estimating the reliability of school means for mixed effects and fixed effects models. For example, Formula (20) can also be viewed as the reliability of school means when items are random and persons are fixed.

In this paper we have concentrated on a relatively straight-forward extension of Hoyt's randomized block analysis of variance technique for calculating reliability; however, the split-plot design, in even its simplest form, appears to be quite powerful, and it is probably sophisticated enough for most small-scale reliability studies and possibly for many large-scale ones. In addition, the split-plot design is conceptually simple, and variance components and reliability (or generalizability) coefficients are often quite easy to calculate, even without a computer.

Clearly, the ideas presented here might be extended for more complex reliability studies, in which persons are nested within more than



one higher order dimension and/or items are administered under different sets of conditions. The reader interested in such extensions (especially the latter one) should consult Cronbach et al.'s (1972) excellent monograph on the Theory of Generalizability, which provides a framework for considering very sophisticated reliability analyses.

## REFERENCES

- Brennan, R. L. and Kane, M. T. The generalizability of class means. Paper presented at the Annual Meeting of the American Educational Research Association, Washington, D.C., April, 1975.
- Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, 16, 297-334.
- Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. *The dependability of behavioral measurements*. New York: Wiley, 1972.
- Cronbach, L. J., Rajaratnam, N., and Gleser, G. C. Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, 1963, 16, 137-163.
- Hoyt C. Test reliability estimated by analysis of variance. *Psychometrika*, 1941, 6, 153-160.
- Kirk, R. E. *Experimental design: Procedures for the behavioral sciences*. Belmont, California: Wadsworth, 1968.
- Kuder, G. F. and Richardson, M. W. The theory of the estimation of test reliability. *Psychometrika*, 1937, 2, 151-160.
- Lindquist, E. F. *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin, 1953.
- Millman, J. and Glass, G. V. Rules of thumb for writing the anova table. *Journal of Educational Measurement*, 1967, 4, 41-51.
- Stanley, J. C. Reliability. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington: American Council on Education, 1971. Pp. 356-442.

## SOME COMMENTS CONCERNING THE USE OF MONOTONIC TRANSFORMATIONS TO REMOVE THE INTERACTION IN TWO-FACTOR ANOVA'S

SCHUYLER W. HUCK AND CARY O. SUTTON

The University of Tennessee

In his discussion of two-factor experiments, Winer (1971) points out that it may be desirable to remove the interaction (and thus obtain additivity of effects) through a monotonic data transformation. The present authors extend Lubin's (1961) discussion of ordinal and disordinal interactions by introducing the concept of "dual-ordinal." This concept is important since a transformation cannot bring about additivity of effects unless the interaction is "dual-ordinal" in nature. For the applied researcher, a simple rule-of-thumb is set forth which allows one to determine, through visual inspection of a *single* interaction graph, whether or not an interaction is dual-ordinal.

When graphed, a significant interaction from a two-factor analysis of variance can take one of two forms. Using Lubin's (1961) terminology, the interaction will either be "ordinal" or "disordinal." If the lines in the graph do not cross one another, the interaction is ordinal. If, however, two or more of the lines cross, the interaction is disordinal. Although his terms are used less frequently, Kerlinger (1964) distinguishes between these two forms of interaction by means of the adjectives "non-symmetric" and "symmetric," with the former term descriptive of a graph in which the lines do not cross.

In his discussion of factorial experiments, Winer (1971) explains that it may be desirable to remove the interaction (and thus obtain additivity of effects) through the use of a monotonic data transformation. As Winer points out, however,

If the means for the levels of factor A have the same rank for all levels of factor B, then a monotonic transformation can potential-

ly remove the  $A \times B$  interaction. When such rank order is not present, a monotonic transformation cannot remove the  $A \times B$  interaction (p. 399).

In other words, a transformation cannot be used to "get rid of" an interaction if that interaction is disordinal. If the interaction is ordinal, however, a transformation may be able to eliminate it from both the data and the underlying model. Hence, the distinction between ordinal and disordinal interactions has a definite practical implication for the applied researcher.

Unfortunately, whether or not an interaction is ordinal can depend upon the way the researcher constructs his graph. While the ordinate is always labelled with the dependent variable, the abscissa can be labelled with the levels of *either* factor A or factor B. If the interaction were to be graphed both ways (i.e., first with factor A on the abscissa, and then again with factor B on the abscissa), it is possible that both graphs might reveal an ordinal interaction or that both graphs might reveal a disordinal interaction. It is also possible, however, for one of the graphs to reveal an ordinal interaction while the other graph reveals a disordinal interaction. To verify this latter possibility, consider the following set of hypothetical cell means for a  $2 \times 2$  design:

		FACTOR B	
FACTOR A	5	10	
	20	15	

When the interaction is graphed with factor B on the abscissa, the two lines do not cross. However, when re-graphed with factor A on the abscissa, the two lines do, in fact, cross one another.

A data transformation can potentially remove the interaction only when the interaction turns out to be ordinal when graphed both ways. In other words, the interaction must be "dual-ordinal" in order for the transformation to have a chance of bringing about a condition of additivity. If either or both of the two graphs turn out to be disordinal, it will be impossible for the transformation to remove the interaction.

Most researchers graph an interaction just once, with the abscissa labelled with the factor that makes the most intuitive sense within the context of the experiment. If this initial (and only) graph reveals a disordinal interaction, then the researcher will immediately know that his interaction is not dual-ordinal. On the other hand, if this graph turns out to be ordinal, the interaction may or may not be dual-ordinal. The

following rule-of-thumb will allow the researcher to determine which is the case (without his having to construct a second graph):

If all of the lines in the first graph slope upward, or if they all slope downward, then the interaction, if re-graphed with the other factor on the abscissa, would also be ordinal. Thus, the interaction is dual-ordinal. Conversely, if one or more of the lines in the initial graph have an upward slope while one or more of the other lines have a downward slope, then the interaction, if re-graphed the other way, would be disordinal. In this case, the interaction is obviously not dual-ordinal.

It should be noted that when an interaction is dual-ordinal in nature, a monotonic transformation *may* bring about a condition of additivity. Removal of the interaction, however, is not guaranteed. In some instances, the original interaction will vanish after the raw scores are subjected to a data transformation. In other instances, the interaction will continue to exist no matter what type of transformation is employed. For this reason, the rule-of-thumb presented above should prove to be more helpful in identifying interactions that *cannot* be eliminated via transformation than in identifying those that can be. In a sense, therefore, the examination of a graphed interaction in light of the "dual-ordinal" concept should be looked upon as a useful screening technique which will prevent the researcher from wasting time trying to eliminate an interaction through a transformation when it is simply impossible to do so.

## REFERENCES

- Kerlinger, F. N. *Foundations of behavioral research*. Holt, Reinhart and Winston, New York, 1964.
- Lubin, A. The interpretation of significant interaction. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1961, 21, 807-817.
- Winer, B. J. *Statistical principles in experimental design*. McGraw-Hill, New York, 1971.





## COMPARING THE VARIANCES OF SEVERAL TREATMENTS WITH A CONTROL

KENNETH J. LEVY

State University of New York at Buffalo

The Dunnett procedure for comparing several treatment means with a control is applied to the problem of comparing several treatment variances with the variance of a control. Appropriate critical values are specified and an example is provided.

OFTEN times, the experimenter is interested in making inferences about treatment variances instead of, or in addition to, inferences about means. Levy (1975) has proposed three multiple range tests for variances which allow pairwise comparisons among  $k$  independent sample variances; the philosophy underlying these tests with respect to the probability of a type-I error is that advocated by Newman (1939) and later by Keuls (1952). The third test suggested by Levy is based upon a normalizing log transformation of the sample variances introduced much earlier by Bartlett and Kendall (1946). If an experimenter were specifically interested in making pairwise comparisons of the variances of several treatments with a control, this same normalizing log transformation could serve as a basis for such a procedure.

Dunnett (1955, 1964) addresses himself to the problem of comparing a set of  $k-1$  treatment means with the mean of a control condition. Each of the resulting  $k-1$  comparisons utilizes the same information concerning the control condition; thus, the  $k-1$  comparisons are not independent. Rather than setting a level of significance equal to  $\alpha$  for each of the individual comparisons, Dunnett establishes a joint significance level for the set of all  $k-1$  comparisons. In the present paper, this same procedure is applied to the problem of comparing several independent treatment variances with the variance of an independent control.

*Theoretical Basis*

Bartlett and Kendall (1946) investigated a normalizing log transformation of the sample variance  $s^2$  when sampling from a  $N(\mu, \sigma^2)$  population. They showed that  $\log_e s^2$  is approximately normally distributed as  $N(\log_e \sigma^2, 2/n)$  where  $n$  is the number of degrees of freedom for  $s^2$ .

Let  $s_0^2, s_1^2, s_2^2, \dots, s_{k-1}^2$  be  $k$  independent sample variances each based upon random samples of size  $n + 1$  drawn from  $k$  independent normal populations,  $N(\mu_i, \sigma_i^2)$ , with unknown means and variances. Let  $l_0, l_1, l_2, \dots, l_{k-1}$  be the log transformations of the given sample variances; i.e.

$$l_0 = \log_e s_0^2.$$

Then, the  $l_i$  are independently distributed approximately as

$$N(\mu_{1i}, \sigma_{1i}^2)$$

where  $\mu_{1i} = \log_e \sigma_i^2$  and  $\sigma_{1i}^2 = 2/n$ .

Let us now consider the quantities,  $z_i$  where

$$z_i = \frac{(l_i - l_0) - (\mu_{1i} - \mu_{10})}{2\sqrt{1/n}} \quad i = 1, 2, \dots, k - 1.$$

As Dunnett (1955) points out, lower confidence limits with joint confidence coefficient  $1 - \alpha$  for the  $k - 1$  comparisons  $\mu_i - \mu_0$  will be given by

$$(l_i - l_0) - d_i' 2\sqrt{1/n}, \quad (i = 1, 2, \dots, k - 1),$$

if the  $k - 1$  constants  $d_i'$  are chosen such that

$$P(z_1 < d_1', z_2 < d_2', \dots, z_{k-1} < d_{k-1}') = 1 - \alpha.$$

Similarly, upper confidence limits will be given by

$$(l_i - l_0) + d_i' 2\sqrt{1/n};$$

and, two-sided confidence limits having the desired joint confidence coefficient will be given by

$$(l_i - l_0) \pm d_i'' 2\sqrt{1/n} \quad (i = 1, 2, \dots, k - 1),$$

if the  $k - 1$  constants  $d_i''$  are chosen to satisfy

$$P(|z_1| < d_1'', |z_2| < d_2'', \dots, |z_{k-1}| < d_{k-1}'') = 1 - \alpha.$$

To find any set of constants  $d_i'$  or  $d_i''$  satisfying these equations, the joint distribution of the  $z_i$  is required. This distribution is a multivariate normal distribution with means 0 and variances 1, where the correlation between  $z_i$  and  $z_j$  is  $1/2$ . Tabulations, for equal values of the

arguments, may be obtained from Dunnett's original one-sided  $t$  tables and Dunnett's revised two-sided  $t$  tables with degrees of freedom equal to  $\infty$ . Thus, for one-sided comparisons between  $k - 1$  variances and a standard variance for a joint confidence coefficient of .95 or .99, the critical values are:

$1 - \alpha$	$k - 1$ , number of variances (excluding the standard)								
	1	2	3	4	5	6	7	8	9
.95	1.64	1.92	2.06	2.16	2.23	2.29	2.34	2.38	2.42
.99	2.33	2.56	2.68	2.77	2.84	2.89	2.93	2.97	3.00

For two-sided comparisons between  $k - 1$  variances and a standard variance for a joint confidence coefficient of .95 or .99, the critical values are:

$1 - \alpha$	$k - 1$ , number of variances (excluding the standard)								
	1	2	3	4	5	6	7	8	9
.95	1.96	2.21	2.35	2.44	2.51	2.57	2.61	2.65	2.69
.99	2.58	2.79	2.92	3.00	3.06	3.11	3.15	3.19	3.22

### Example

Suppose that a completely randomized experiment with three treatments and a control has been conducted. Suppose further, that the experiment employed 10 subjects per treatment and that the experimenter is specifically interested in comparing the variances of each of the three treatments with the variance of the control; perhaps mean differences are unimportant or perhaps mean differences are not predicted at all.

Suppose that the following results were obtained:

	$s^2$	$\log_e s^2$
Control	12.00	2.4849
Trt 1	.75	-.2877
Trt 2	3.00	1.0986
Trt 3	2.00	.6931

where

$$s_i^2 = \sum_{j=1}^n \frac{(x_{ij} - \bar{x}_i)^2}{n - 1},$$

$n$  = the number of subjects in the  $i$ th group, and  $n - 1$  = degrees of freedom for  $s_i^2$ .

Let us now test to see whether  $\sigma_1^2$ ,  $\sigma_2^2$ , and  $\sigma_3^2$  differ significantly from  $\sigma_0^2$ , i.e., we wish to test the null hypothesis

$$H_0: \sigma_0^2 = \sigma_i^2$$

against the alternative

$$H_1: \sigma_o^2 \neq \sigma_i^2$$

for  $i = 1, 2, 3$ . When the null hypotheses are true,

$$z_1 = \frac{(-.2877) - (2.4849)}{2\sqrt{1/9}} = -4.1589$$

$$z_2 = \frac{(1.0986) - (2.4849)}{2\sqrt{1/9}} = -2.0794$$

$$z_3 = \frac{(.6931) - (2.4849)}{2\sqrt{1/9}} = -2.6877$$

For an overall  $\alpha$  level controlled at .05, one would reject the hypotheses  $H_0: \sigma_o^2 = \sigma_1^2$  and  $H_0: \sigma_o^2 = \sigma_3^2$  since  $-4.1589$  and  $-2.6877$  are both less than  $-2.35$ .

## REFERENCES

- Bartlett, M. S. and Kendall, D. G. The statistical analysis of variance heterogeneity and the logarithmic transformation. *Royal Statistical Society*, 1946, 8, 128-138.
- Dunnett, C. W. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 1955, 50, 1096-1121.
- Dunnett, C. W. New tables for multiple comparisons with a control. *Biometrics*, 1964, 20, 482-491.
- Keuls, M. The use of the "studentized range" in connection with an analysis of variance. *Euphytica*, 1952, 112-122.
- Levy, K. J. Some multiple range tests for variances. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1975, 35, 599-604.
- Newman, D. The distribution of the range in samples from a normal population, expressed in terms of an independent estimate of standard deviation. *Biometrika*, 1939, 31, 20-30.

## NEGATIVE SIMILARITIES<sup>1</sup>

ULF LUNDBERG AND BERNARD DEVINE

University of Stockholm, Sweden

Two experiments were carried out for the present investigation. The first experiment was an exact replication of an experiment by Ekman (1955), where the subjects had been requested to estimate positive similarity between pairs of emotional terms. The second experiment was carried out in the same way except that the subjects were also requested to give negative estimations when they considered that a pair of words described feelings which were opposite to each other. Using factor analysis it was found that the negative estimations obtained in the second experiment were represented as zero ratings in the first one. The second experiment also yielded some additional information which was considered to be psychologically meaningful. A reanalysis of Ekman's data (1955) gave almost exactly the same result as the first experiment in the present study.

IN the early 1960's Ekman developed a multidimensional scaling model in which the similarity between two stimuli is directly represented by the angle between and the length of hypothetical vectors (Ekman, 1963; Ekman, Engen, Künnapas and Lindman, 1964), these being used to represent the qualitative and quantitative difference between the stimuli, respectively. For the case where all the vectors were assumed to be of equal length the formula found to fit the empirical data was

$$s_{ij} = \frac{\cos \varphi_{ij}}{\cos (\varphi_{ii}/2)}, \quad (1)$$

where  $s_{ij}$  represents the similarity estimate and  $\varphi_{ij}$  the angular separation between stimulus  $i$  and  $j$ . Later Ekman (1965) presented a refor-

---

<sup>1</sup> The research was supported by grants from the University of Stockholm and from the Swedish Council for Social Science Research.



mulation of the above equation solved for cosine  $\varphi_{ij}$

$$\cos \varphi_{ij} = (s_{ij}/4) (s_{ij} + \sqrt{8 + s_{ij}^2}), \quad (2)$$

and he noted that when the similarities are given on a scale from 0 to 1, the cosines derivable from the relation above differ from the actual similarity estimates by 0.07 at the most and, for practical purposes, the actual similarities may be used in the analysis, as had been the custom before the model was developed (Ekman, 1954, 1955). However, the model is only meaningful for positive values, i.e., when the angle between the vectors is less than  $90^\circ$ . Therefore, when the model is applied the subjects are requested to give similarity estimations that are either positive or zero. Micko (1970) has produced a modification of Ekman's model which he terms a 'halo'-model. Negative scalar products of the stimulus vectors are possible within this modification.

An alternative method of analysis was tried by Stone and Coles (1970), who used a technique where they correlated the columns of a similarity matrix and constructed a new matrix from the correlations. The matrix obtained in this manner also includes negative values. When they reanalyzed the data gained by Ekman (1965) and Künnapas (1966), they found that a factor analysis yielded a smaller number of factors than was found in the original studies. However, although all the new factors were bipolar while the original ones were all unipolar, no new information appeared. A comparison between the Micko and Stone-Coles approach was made by Stone (1971).

Although the expression "negative similarities" may itself sound contradictory, it does not follow that subjects would not be able to produce meaningful results were they to be given the possibility of using both positive and negative estimations. One area in which this possibility does not seem too unrealistic is in the judgment of similarity between emotional terms, where some words describe feelings which may be perceived as positively related, e.g., "glad-happy" while others appear more or less opposite to each other, e.g., "glad-sad." The psychological meaningfulness of opposing stimuli of this kind has been pointed out by Stone and Coles (1970) and by Yoshida, Kinase, Kurokawa, and Yashiro (1970). Recently a study by Tucker (1972) has been published where use is made of this bipolar relation in order to demonstrate a development of three-mode factor analysis. For the present report the opposing relation has been termed "negative similarity."

The present experiment was conducted in order to study the relation between the results from negative similarity estimations and the results gained from studies where only positive estimations had been used.

### *Method*

#### *The Pilot Study*

In 1955 Ekman conducted a similarity study of emotional terms. He analyzed the data using Thurstone's centroid method (1947) and found eleven factors. Fifteen of the 23 words used in his study were used in a pilot study for the present investigations, where the aim was to find out if a sufficient number of the words were consistently perceived as being of an opposite nature to each other. The 15 words were paired in all the possible combinations and presented to 17 psychology students. The subjects were instructed to put a plus sign if they considered that the pair described similar feelings and a minus sign for opposing feelings.

It was found that 25 of the 105 pairs were perceived as opposites by every one of the subjects, whereas only five of the pairs were consistently perceived as similar. Consequently, it was considered possible to use these stimuli for studying negative similarities.

#### *The Main Experiments*

Ekman (1955) used 168 psychology students in his study. The subjects were requested to make their estimations on a 5-point scale ranging from 0 to 4, and to judge the qualitative similarity between the words describing emotional states. Zero indicated "no similarity at all" and four indicated "identity." It was assumed that the emphasis on the qualitative similarity would result in estimates of equally intense emotions which could be represented by vectors of equal length in the multidimensional model. Thus after a transformation to a scale from 0 to 1, the similarity estimates should not deviate very much from the theoretical cosine values (see also Ekman, 1970).

Two main experiments were carried out for the present investigation. As far as possible Experiment I was an exact replication of Ekman's study from 1955, and 150 psychology students were used as subjects. The purpose of this experiment was to yield up-to-date results concerning the semantic meaning of the emotional terms. This was considered necessary as some changes in the usage of language may have occurred during this period. Yoshida, Kinase, Kurokawa and Yashiro (1970) take up some aspects of this point. Another 150 students took part in Experiment II, and the procedure was essentially the same as in Experiment I. The only difference was that the subjects in the second study were requested to give similarity estimations on a nine point scale ranging from +4 to -4, where positive values in-

licated the degree of positive similarity between the feelings described by the words, negative values indicated the degree of opposite relationship, and zero indicated that the words were in no way related to each other (see also Table 2).

The English words presented by Ekman (1955) and referred to in the present investigation may in some cases deviate from the translations which we consider to give a better interpretation of the Swedish words. For these cases we have put Ekman's translations within parenthesis. One Swedish word (*saknad*) seemed to require the two English words "miss" and "lack" to clarify its meaning.

### *Results*

The arithmetic means of the estimations were calculated for each pair of words in both experiments, and a uniform reduction of the scale was made to a range from 0 to +1, and from -1 to +1, respectively. A principal component solution was performed on the two matrices. Unity was used in the main diagonal for the communality estimations, and seven factors were extracted and varimax rotated for Experiment I and six for Experiment II. The number of factors extracted was decided according to the number of eigenvalues greater than unity.

The factors obtained in Experiments I and II were compared to each other and interpreted as shown in Table 1. It was found that all the factors obtained in Experiment I correspond closely to factors from Experiment II in the way schematically demonstrated in Figure 1. Four factors from Experiment II appeared as bipolar factors and two as unipolar factors. There is one bipolar factor in Experiment II (Discontentedness and Contentedness), which did not appear in Experiment I, and the suggestion of the second half of a bipolar factor (Passive Repulsion). Not too much emphasis should be placed on the particular labels used to denote the factors. Nevertheless, it is clear that all the information gained Experiment I was available in Experiment II, and moreover, some additional information was obtained.

In order to investigate the manner in which the results from Experiments I and II are related to each other, three hypotheses were suggested.

Hypothesis 1.—The negative values in Experiment II were represented as zero ratings in Experiment I. The implication is that the subjects were able to discriminate between the zero ratings in Experiment I but had no means available to represent this differentiation.

Hypothesis 2.—The negative values in Experiment II were represented as the lower half of the positive scale in Experiment I. This

TABLE I  
Factor Loadings > |0.30| from Experiments I and II

A. Depression and Elation			B. Active Attraction and Active Repulsion		
Word	Exp I	Exp II	Word	Exp I	Exp II
	Factor I	Factor I		Factor VI	Factor II
Sad	0.94	0.98	Loving (Affectionate)	0.89	0.90
In low spirits (Depressed)	0.90	0.91	Tenderness	0.87	0.87
Desperate	0.73	0.74	Benevolent	0.71	0.59
Lack, Miss (Want)	0.55	0.52	Happy	0.63	0.58
Anxious	0.31	0.40	Glad	0.43	0.35
	Factor II			Factor III	
Gay	0.92	-0.82	Hatred (Rancorous)	0.90	-0.93
Glad	0.86	-0.83	Detest (Disgust)	0.87	-0.80
Animated	0.79	-0.53	Rage (Ireful)	0.79	-0.52
Happy	0.59	-0.76	Vexed (Angry)	0.65	
			Irritated	0.41	

TABLE 1 (Continued)

C. General Agitation			D. Passive Attraction and Passive Repulsion		
Word	Exp I	Exp II	Word	Exp I	Exp II
	Factor III			Factor V	
Impatient	0.88	0.79	Longing	0.90	0.88
Irritated	0.79	0.52	Desire	0.84	0.84
Restless	0.77	0.91	Lack, Miss (Want)	0.69	0.66
Vexed (Angry)	0.54				
Agitated	0.51	0.75	Detest (Disgust)		-0.38
Rage (Ireful)	0.40				
Anxious	0.31	0.56			
Animated		0.49			
E. Discontentedness and Contentedness					
	Factor V		F. Fear		
Word	Exp I	Exp II	Word	Exp I	Exp II
	Factor VI			Factor VII	
Vexed	0.84		Anxious	0.81	-0.58
Rage (Ireful)	0.73		Frightened	0.79	-0.90
Irritated	0.67		Agitated	0.72	-0.50
Impatient	0.42		Desperate	0.38	
Gay		-0.51	Restless	0.37	
Benevolent		-0.47			
Glad		-0.45			



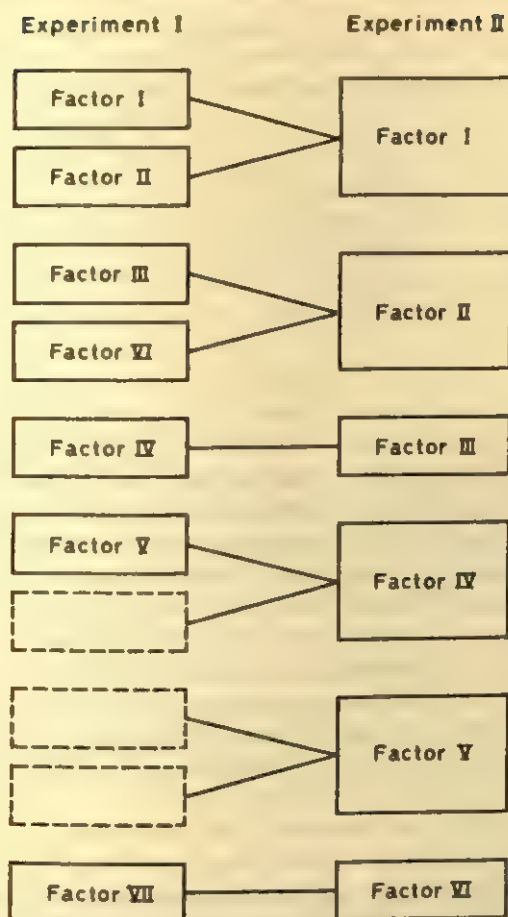


Figure 1. The diagram shows the corresponding factors obtained in Experiments I and II. The broken lines indicate that no corresponding factor was obtained in Experiment I.

implies that the subjects spread out the positive values so that the lower values were shifted onto the negative side.

Hypothesis 3.—The negative values in Experiment II were a reflection of some of the positive values given in Experiment I. This implies that if the negative values are pivoted around the zero point of the scale, then Experiment II will give the same results as Experiment I.

Table 2 shows the percentage distributions for the different scale values in Experiments I and II, and also the way in which the empirical scale values were transformed according to the three hypotheses. The transformed values were then uniformly reduced to range from 0 to 1 and were factor analyzed by the same method as used above. For

TABLE 2

*Actual and Transformed Similarity Scale Values and Percentage Distribution of Empirical Estimates*

Experiment I	Experiment II			
	Actual Values	Hypothesis 1	Hypothesis 2	Hypothesis 3
+4 ( 3.9 %)	+4 ( 3.5 %)	+4	+4	+4
+3 (10.4 %)	+3 ( 8.9 %)	+3	+3.5	+3
+2 (13.4 %)	+2 (10.9 %)	+2	+3	+2
+1 (17.9 %)	+1 (13.0 %)	+1	+2.5	+1
0 (54.4 %)	0 (37.6 %)	0	+2	0
	-1 ( 4.7 %)	0	+1.5	+1
	-2 ( 6.9 %)	0	+1	+2
	-3 ( 8.2 %)	0	+0.5	+3
	-4 ( 6.3 %)	0	0	+4

Hypothesis 1 seven factors were extracted and rotated to obtain the best comparison with the factor matrix from Experiment I. This was done using a computer program (RELATE) by Veldman (1967). This program rotates a problem factor matrix so that maximum contiguity with a hypothesis (target) matrix is achieved. Tucker's coefficient of congruence (see Harman, 1967, p. 270) was then calculated as a measure of factor similarity and the resulting values are presented in Table 3. All the coefficients are higher than 0.995, which shows that these two matrices are almost exactly the same. The distributions of the estimates shown in Table 2 also indicate that Hypothesis 1 is confirmed.

Three factors were extracted for Hypothesis 2 and five for Hypothesis 3, the number of factors extracted being decided according to the eigen-values, using the same criterion as given above. The two factor matrices were also rotated against the matrix from Experiment I in order to obtain maximum contiguity, and the resulting coefficients of congruence are shown in Table 3. These coefficients are considerably lower than those obtained according to Hypothesis 1. It is clear that Hypotheses 2 and 3 are inferior in describing the data from

TABLE 3  
*Tucker's Coefficient of Congruence*

Comparison	Factor						
	I	II	III	IV	V	VI	VII
Exp. I/Exp. II(Hyp. 1)	0.997	0.999	0.998	0.998	0.996	0.996	0.997
Exp. I/Exp. II(Hyp. 2)	0.917	0.887	0.932	—	—	—	—
Exp. I/Exp. II(Hyp. 3)	0.797	0.609	0.641	0.895	0.958	—	—
Exp. I/Ekman 1955	0.996	0.995	0.998	0.997	0.995	0.993	0.993

Experiment I, both with respect to the number of factors and to their loadings.

A reanalysis of Ekman's data (1955) was made by the present authors using a principal component solution and this analysis suggested seven factors. This solution was rotated against Experiment I and the coefficients of congruence for the comparison are included in Table 3. The coefficients show that there is a very strong relationship. Earlier reanalyses of Ekman's data have been carried out by Dietze (1963) and Waern (1972), however, a comparison with their results is difficult as they used two kinds of typal analyses.

### *Discussion*

The results from this investigation show the relation between an experiment with only positive estimations (Experiment I) and one that included negative similarity estimations (Experiment II). The negative similarities were found to have been represented mainly as zero ratings when only a positive scale was available for the subjects. Two questions arise naturally from this finding. (1) Do the negative similarity estimations from this study contain any meaningful information? (2) Does the relation obtained in this study also hold for other kinds of stimuli?

There are two reasons why the results from the present experiment indicate an affirmative answer to the first question. The first reason is that the negative similarities have yielded bipolar factors which seem to be psychologically meaningful. The second reason is that the negative similarities yielded more information in the analysis than was obtained when only positive estimations were used. The bipolarity was, of course, suggested by the scaling method used, however, this was also the case for the unipolar factors obtained when positive similarity estimations were used in other studies.

It is interesting to note that two factors appeared that were not bipolar, these being interpreted as Fear and General Agitation. There are at least two possibilities why these factors have turned out to be unipolar. The first one is that the words which might appear on the other half of a bipolar factor were not included among the stimuli presented to the subjects. The sample of stimuli used in a study like this one is very important for both the number of factors obtained and the interpretations that can be made of them. A second possibility is that the factors represent emotions which are genuinely unipolar. In the case of Fear it is not easy to find an obvious opposite to the word regarding the emotional context corresponding to its meaning, for it does not seem unreasonable to think of Fear as starting from some

kind of "zero" point. In the case of General Agitation it is quite possible that words like "calm" are experienced as psychologically opposite to "restless" within the context of the (cognitive) situation, nevertheless, "calmness" would only seem to represent a lesser degree of "restlessness" (activity related to a state of balance).

The bipolarity of another of the factors may be questionable. This factor was interpreted as Attraction and Repulsion (passive) and it has only one moderate loading on the Repulsion side (0.38 for "detest" [disgust]). In this case we believe that the lack of words on this side are more likely to be due to the sample of emotional terms used rather than the lack of real bipolarity for this factor. Psychologically, the bipolarity seems meaningful within the interpretations given above.

The additional information obtained by using negative similarities was one complete bipolar factor and the negative side of another bipolar factor. The half of the bipolar factor was named Passive Repulsion and has already been commented on. The complete bipolar factor was interpreted as Discontentedness and Contentedness, and the words highly loaded on each side are "vexed" (angry), "rage" (ireful), and "gay," "benevolent," respectively. These words were included in other factors in Experiment I, particularly "vexed" (angry) and "rage" (ireful) which also appeared in Active Repulsion. We are tempted to describe the qualitative nature of Active Repulsion and Discontentedness as, in the former case, a feeling directed towards a being, and in the latter case, a general condition having more to do with a lack of emotional balance and harmony. The second half of the latter factor, Contentedness, seems to be less definite as there are no very high loadings which would make the interpretation unambiguous. This may also be due to the particular sample of words used.

The second question mentioned above can only be answered by a study using other stimuli.

Finally, it should be remembered that both the present study and Ekman's (1955) were carried out with Swedish subjects and with Swedish words, and that the interpretations of the factors were made with the nuances of the Swedish words in mind.

## REFERENCES

- Dietze, A. G. Types of emotions or dimensions of emotions? A comparison of typal analysis with factor analysis. *Journal of Psychology*, 1963, 56, 143-159.
- Ekman, G. Dimensions of color vision. *Journal of Psychology*, 1954, 38, 467-474.
- Ekman, G. Dimensions of emotion. *Acta Psychologica*, 1955, 11, 279-288.

- Ekman, G. A direct method for multidimensional ratio scaling. *Psychometrika*, 1963, 28, 33-41.
- Ekman, G. Two methods for the analysis of perceptual dimensionality. *Perceptual and Motor Skills*, 1965, 20, 557-575.
- Ekman, G. Comparative studies on multidimensional scaling and related techniques. *Reports from the Psychological Laboratories, The University of Stockholm*, 1970, Suppl. 3.
- Ekman, G., Engen, T., Künnapas, T., and Lindman, R. A quantitative principle of qualitative similarity. *Journal of Experimental Psychology*, 1964, 68, 530-536.
- Harman, H. H. *Modern factor analysis*. Chicago: University of Chicago Press, 1967.
- Künnapas, T. Visual perception of capital letters; multidimensional ratio scaling and multidimensional similarity. *Scandinavian Journal of Psychology*, 1966, 7, 189-196.
- Micko, H. C. A "halo"-model for multidimensional ratio scaling. *Psychometrika*, 1970, 35, 199-227.
- Stone, L.A. Congruent multidimensional scaling results obtained using the halo-model and the Stone-Coles method-model. *Perceptual and Motor Skills*, 1971, 33, 524-526.
- Stone, L. A. and Coles, G. J. Correlational similarity: The basis for a new revised method of similarity analysis. *Studia Psychologica* (Bratislava), 1970, 12, 258-265.
- Thurstone, L. L. *Multiple-factor analysis*. Chicago: University of Chicago Press, 1947.
- Tucker, L. R. Relations between multidimensional scaling and three-mode factor analysis. *Psychometrika*, 1972, 37, 3-27.
- Veldman, D. J. *FORTRAN programming for the behavioral sciences*. New York: Holt, 1967.
- Waern, Y. Structure in similarity matrices. A graphic approach. *Scandinavian Journal of Psychology*, 1972, 13, 5-16.
- Yoshida, M., Kinase, R., Kurokawa, J., and Yashiro, S. Multidimensional scaling of emotion. *Japanese Psychological Research*, 1970, 12, 45-61.





## SCALING ATTITUDE ITEMS: A COMPARISON OF SCALOGRAM ANALYSIS AND ORDERING THEORY

PETER W. AIRASIAN, GEORGE F. MADAUS, AND ELINOR M. WOODS

Boston College

The study compared an ordering-theoretic method of identifying item hierarchies with scalogram analysis in the evaluation of an eight item attitude measure. The attitude measure assessed "progressive" and "traditional" views of education. Data were collected in a survey of a random sample of 178 parents of public school children. The scalogram analysis revealed that the items did not form a unidimensional and cumulative hierarchy. The ordering-theoretic analysis identified a branched, nonlinear hierarchy which had higher reproducibility and scalability than the linear hierarchy identified by the scalogram analysis. The results support the use of ordering theory in defining item hierarchies in attitudinal measures.

THE purpose of this study was to compare the results of a scalogram analysis (Guttman, 1944, 1950) of eight items assessing attitude toward education to the results of an ordering theoretic analysis, a mode of analysis which is capable of identifying branched hierarchies among items (Airasian and Bart, 1973). Utilizing statistics such as reproducibility (Guttman, 1950) and the coefficient of scalability (Anderson, 1966), the study sought to identify which method of analysis provided an item hierarchy which best fitted the observed data.

The classical method of scaling attitudinal items into an ordered hierarchy is scalogram analysis (Guttman, 1944, 1950). Scalogram analysis is used to order a group of items into a linear hierarchy and to evaluate whether or not the hierarchy is unidimensional and cumulative. The degree to which a group of items is judged to possess these properties is determined by the extent to which "passes" (scores of 1) on any item co-occur with "passes" on all items ranked lower in

the hierarchy. The inverse is also true. That is, a scale is unidimensional and cumulative insofar as "failures" (scores of 0) on an item co-occur with "failures" on items ranked higher in the hierarchy.

Except in a few rare instances, most notably the articulation of social distance scales, scalogram analysis has been used with disappointing results. A primary reason for the lack of demonstrated viability of the approach is the constraint that scaled items must manifest a linear hierarchy. A linear hierarchy is one in which one and only one item appears at any given hierarchic level and one and only one item is immediately prerequisite to or a sufficient condition for any given item in the hierarchy. As a consequence of this constraint, reliable Guttman scales exceeding five items are rare. Methodological extensions of scalogram procedures such as multiple scalogram analysis (Lingoes, 1963) and latent structure analysis (Lazarsfeld, 1959) are similarly constrained to identifying only linear hierarchies among items.

In addition to methodological difficulties in defining reliable linear hierarchies consisting of more than five items, there are conceptual problems associated with the search for item domains which can appropriately be represented by linear hierarchies. Recent research in the areas of cognitive development (e.g., Airasian and Bart, 1972), tests and measurement (e.g., Airasian, 1971a; Cox and Graham, 1966, Walbesser, 1968), and curriculum development (e.g., Gagne, 1962, 1967, 1968; Gagne and Paradise, 1961) has demonstrated that groups of test items or curricular tasks are more likely to be ordered in nonlinear, branched hierarchies than in linear hierarchies. While nonlinear hierarchies do not manifest the conceptual simplicity of linear hierarchies, it is likely that for most sets of test items or curricular tasks, nonlinear hierarchies represent a more accurate description of the relationships between items or tasks than do linear hierarchies.

### *Ordering Theory*

Ordering theory (Airasian and Bart, 1973; Bart and Krus, 1973) is a measurement model which is based upon scalogram analysis but which extends scalogram techniques to the articulation of nonlinear item or task hierarchies. Ordering theory is an approach to fundamental measurement and has as its primary intent either the determination of a hierarchy for a set of items or tasks or the verification of an *a priori*, hypothesized hierarchy among a set of items and tasks. At a foundational level of explanation, ordering analysis possesses a boolean algebraic framework in which item or task response patterns are viewed as atoms in a boolean algebra with as many generators as there are items being considered (Goodstein, 1963). Although there

are many ways to describe ordering theory, it is most meaningfully classified as a deterministic measurement model which uses item response patterns as the basic data points to identify both linear and nonlinear, qualitative, prerequisite relations among test items or tasks. In this regard, it is a more general case of scalogram analysis.

Ordering-theoretic procedures were used in this study to determine prerequisite relations between pairs of items assessing attitude toward education. The prerequisite relation is considered here, since that type of relation is of primary interest to behavioral scientists, especially in their quest to identify sequential or causal relationships among phenomena.

An item  $i$  is a prerequisite to an item  $j$  to the extent that the (0, 1) response pattern, where 0 represents the score on item  $i$  and 1 represents the score on item  $j$ , occurs infrequently. The (0, 1) response pattern is viewed as a disconfirmation that a correct or acceptable response to item  $i$  is a prerequisite to a correct or acceptable response to item  $j$ . For any pair of items, a  $2 \times 2$  table showing the number of "passes" or "fails" on the items can be constructed. Table 1 represents a hypothetical example of such a table. In Table 1, the (0, 1) response cell has a frequency of 0, indicating that no subject attained item  $j$  after failing item  $i$ . As a consequence, item  $i$  can be considered to be prerequisite to item  $j$ . Note that if the (1, 0) cell also had a frequency of 0, so that all subjects manifested either a (0, 0) or (1, 1) response pattern, item  $i$  and  $j$  would be equivalent. That is, they would not be hierarchically ordered and would, in fact, be extracting redundant information. Constructing contingency tables such as shown in Table 1 for all possible item pairs is one method of extracting the prerequisite relations among a set of items. In this study, a computer program (Lele and Bart, 1971) was utilized in the data analyses.

In defining prerequisite relations between items, ordering theory shares one limitation with scalogram analysis. Both measurement models are deterministic. Thus, neither incorporates a method of dealing with the possibility of encountering random error in item response patterns. To overcome this limitation, ordering theoretic analyses rely upon the use of a preset tolerance level for error. The tolerance level sets the number of disconfirmatory response patterns which will be accepted in defining a prerequisite relation between two items. Thus, for

TABLE 1  
*Example of a  $2 \times 2$  Table to Determine a Prerequisite Relation between Two Items*

		item $j$	
		0	1
item $i$	0	40	0
	1	40	20

a 5% tolerance level and  $N$  subjects, one would tolerate at most  $(.05)N$  disconfirmatory response patterns between items in an item pair before accepting a prerequisite relation. Referring to Table 1, for a 5% tolerance level and 100 subjects, 5 (0, 1) response patterns could be manifested before the prerequisite relation between items  $i$  and  $j$  was rejected.

There are two general strategies for the implementation of tolerance levels in examining the hierarchical structure or ordering among a set of items. Within one strategy, a hierarchy and its array of prerequisite relations is hypothesized for a set of items or tasks before data collection. The response patterns that would disconfirm the entire hierarchy are identified and a tolerance level is established (Airasian, 1971b, 1971c). The hypothesized hierarchy is then accepted if the frequency of obtained disconfirmatory response patterns is less than or equal to the prescribed tolerance level.

An alternative strategy can be used when no a priori hierarchy among items is hypothesized. This strategy, which was used in this study, identifies prerequisite relationships between item pairs from the obtained response patterns. In this approach, all possible item pairs are investigated to identify prerequisite relations. The prerequisite relation for a particular pair of items is accepted if the frequency of obtained disconfirmatory response patterns for the item pair is less than or equal to the frequency of such response patterns established by the tolerance level. This procedure is followed to test each of the possible hypothesized prerequisite relations with the same tolerance level being used for each testing.

Several discussions of ordering theory have been provided. Airasian and Bart (1973) articulated the general nature of ordering theory. Bart and Krus (1973) described an ordering-theoretic technique by which item or task hierarchies could be determined. Krus and Bart (1974) discussed an ordering-theoretic technique to scale items in a multidimensional setting. Airasian, Bart and Greaney (1973) and Bart and Airasian (1972) investigated the hierarchies among Piagetian tasks by means of ordering theory. Two defining properties of ordering theory cited in these discussions are that all test items or tasks to be examined must be dichotomously scored and that all subjects in a sample must respond to all of the items or tasks. In this study an ordering-theoretic analysis was applied to an attitudinal scale assessing attitude towards education.

### *Procedures*

The attitude scale which was studied consisted of eight statements measuring a "progressive" and "traditional" view of education



(Kerlinger, 1967). The eight items were part of an interview instrument administered to a random sample of 183 parents of public school children in a suburb of Boston. The items comprising the scale are listed in Table 2 along with the responses expected for a subject categorized as a "progressive."

Each of the eight statements permitted five response options: agree strongly, agree somewhat, disagree somewhat, disagree strongly, and no opinion. To meet the constraint of ordering theory that item responses be bivalued, the two "agree" response options were collapsed into one category and the two "disagree" response options into a second category. A total of 5 of the original 183 parents selected the "no opinion" option in responding to one or more of the eight items. These respondents were eliminated from the sample reducing the final sample size in the study to 178.

For the purposes of analysis, the eight statements were scored in the direction of the "progressive" view of education, with the "progressive" response to each item being coded 1 (pass) and the "traditional" response 0 (failure). Thus, if an individual disagreed with item 1 in Table 2, he scored 1 on that item, since his response matched the "progressive" key. The selection of the "progressive" scoring key was arbitrary. The need was to establish, for each respondent, an item response pattern comprised of 0's and 1's in the same metric as the other respondents' patterns. Had "traditional" responses been

TABLE 2  
*"Progressive" and "Traditional" Items and "Progressive" Scoring Key*

Item	Progressive Response
1. Teachers should keep in mind that pupils have to be made to work.	D
2. More authority is needed in today's classroom.	D
3. Student participation in the forming of school policy is a privilege granted by the school rather than a matter of student rights.	D
4. Teachers should encourage pupils to study and criticize our own and other economic systems and practices.	A
5. Children need to have more supervision and discipline than they are getting.	D
6. There should be more emphasis on the three R's.	D
7. Schools should be sources of new social ideas.	A
8. Modern schools have too many fads and frills, such as activity programs, driving education, crafts, social services and the like.	D

selected as the scoring key, the hierarchy produced by the analysis would have been identical to the "progressive" results, except in inverted order. The coefficient alpha reliability of the dichotomized items for the sample of 178 respondents was .71.

### Results

Figure 1 shows the linear hierarchy generated by the scalogram analysis. The numbers in the scale correspond to item numbers in Table 2. Item 4, which stated "Teachers should encourage pupils to study and criticize our own and other economic systems and practices" emerged as the most progressive item. The scale ranged upwards to item 3, "Student participation in the forming of school policy is a privilege granted by the school rather than a matter of student rights," which was the least progressive item in the scale.

The coefficient of reproducibility (Guttman, 1950, Torgerson, 1958), which provides an index of the fit of the data to the linear hierarchy resulting from the scalogram analysis, was .84. For an eight item hierarchy, reproducibility should be a minimum of .9 if the hierarchy is to be considered unidimensional and cumulative.

Three other statistics, the minimum marginal reproducibility, the percentage improvement, and the coefficient of scalability, were calculated for the hierarchy identified by the scalogram analysis (Anderson, 1966). The reproducibility of any item can never be less

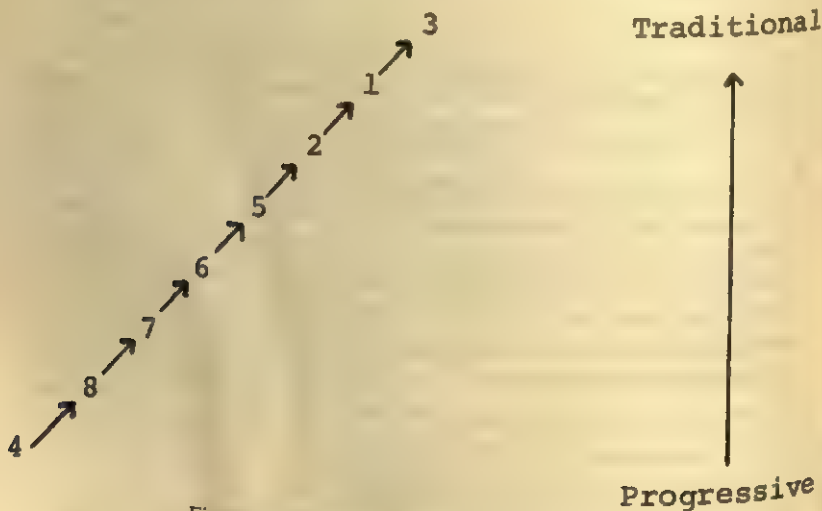


Figure 1. Scalogram Analysis Item Hierarchy.

than the percentage of respondents falling into the most popular response option for that item. The minimum marginal reproducibility is the minimum coefficient of reproducibility that could have occurred for the scale given the percentage of respondents falling into the most popular response option for each item. It is the average across all items of the percentage of respondents selecting the most popular response option on each item. For the scalogram analysis, the minimal marginal reproducibility was .72. The percentage improvement, which indicates the extent to which the coefficient of reproducibility is due to the response patterns and not the percentage of respondents selecting the most popular options on the items, is the difference between the coefficient of reproducibility and the minimal marginal reproducibility. In the case of the scalogram analysis, the percentage improvement was .12, a small value. Finally, the coefficient of scalability, which is obtained by dividing the percentage improvement by 1 minus the minimum marginal reproducibility, was .43 for the scalogram analysis. In order for a hierarchy to be considered truly unidimensional and cumulative, the coefficient should be well above .5 (Anderson, 1966). Overall, then, the scalogram analysis revealed that a linear hierarchy is not an adequate representation of the relationships among the eight items.

Figure 2 shows the hierarchy generated by the ordering-theoretic analysis with a tolerance level set at 10%. Two points about this ordering, relative to the scalogram analysis, are noteworthy. First, the items occupy the same relative positions in both hierarchies, although there are fewer hierarchical levels in the hierarchy generated by the ordering analysis. Thus, while items 1, 2, 3, and 5 are all at the same level in the ordering analysis results, they are at a higher level in the hierarchy than the remaining items. Similarly, item 6 scales above items 8 and 4 in both hierarchies, as does item 8 above item 4. Second, the ordering analysis reveals where the hierarchy for the eight items departs from linearity. Scalogram analysis reveals the best fitting linear hierarchy and the extent to which that hierarchy is unidimensional and cumulatively hierarchical. It cannot reveal departures from linearity. The ordering-theoretic analysis indicated that the best fit for the item response patterns was a non-linear, branched hierarchy. The analysis generated the form of that hierarchy.

Figure 2 indicates prerequisite relations as well as relations of equivalence and logical independence between the eight items. Agreement on one statement is a prerequisite to agreement on another statement if the number representing the first statement is connected to the number representing the second statement by a line that passes in a general upwards direction. For the relation "is a prerequisite to,"

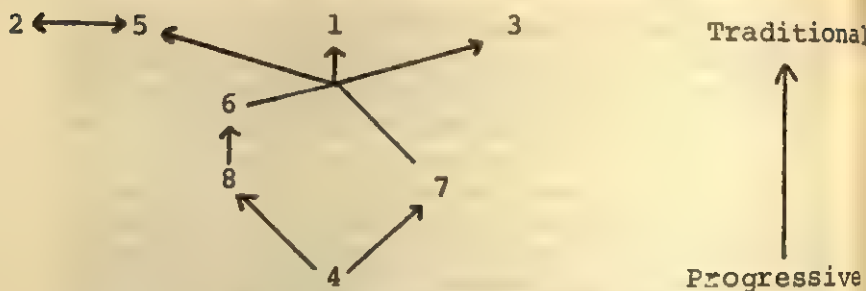


Figure 2. Item Hierarchy from Ordering Analysis.

phrases such as "is a precondition for" and "is a necessary condition for" may be viewed as being synonymous. Also, if success on one task is a necessary condition for success on another task, success on the second task is a sufficient condition for success on the first task. Table 3 lists the logical relations among the eight items identified by the ordering analysis.

While most "progressives" would agree with item 4, the ordering analysis reveals that a split occurs given agreement with item 4. Given that the keyed responses to items 7 and 8 are "agree" and "disagree" respectively, some "progressive" respondents indicated disagreement with item 8, "Modern schools have too many fads and frills . . ." and disagreement with item 7, "Schools should be the source of new social ideas" while others agreed with 8 and 7. The responses of the sample indicate that the items concerned with "fads and frills" in the schools (item 8) and the school teaching new social ideas (item 7) are not on a unidimensional continuum. The ordering analysis reveals that the issue of schools being the source of new social ideas (item 7) is different in kind than the issue of "fads and frills" in school programs (item 8).

TABLE 3  
*Logical Relations among the Eight Items as Defined by the Ordering Analysis*

Item	Relation	Item(s)
4	is prerequisite to	1, 2, 3, 5, 6, 7, 8
8	is prerequisite to	1, 2, 3, 5, 6
7	is prerequisite to	1, 2, 3, 5
6	is prerequisite to	1, 2, 3, 5
2	is equivalent to	5
8	is independent of	7
6	is independent of	7
1	is independent of	2, 5, 3
3	is independent of	1, 2, 5
2	is independent of	1, 3
5	is independent of	1, 3

Note that item 6, "There should be more emphasis on the three R's" is related to item 8 but not to item 7, further corroboration that the respondents perceive the issue of new social ideas to be distinct from concern over the more formal academic program. The ordering analysis also revealed that items 2, 5, 1, and 3 were the least "progressive" items. However, these items were not hierarchically ordered among themselves. Finally, the double arrow between item 2, "More authority is needed in today's classroom" and item 5, "Children need to have more supervision and discipline than they are getting," indicates that these items are equivalent. That is, the parents responded identically to items 2 and 5. If a respondent agreed with 2, he agreed with 5. If he disagreed with 2, he disagreed with 5. In essence, the ordering analysis revealed that these two items were extracting redundant information and that one, but not both, was needed in the scale.

Table 4 compares the scalogram analysis with the ordering-theoretic analysis in terms of the four statistics used to validate a hierarchy. It is evident that the branched hierarchy identified by the ordering analysis is more reproducible than the linear scalogram analysis hierarchy. The nonlinear hierarchy evidenced a reproducibility well above the lower limit of scalability discussed by Torgerson (1958). Since the same item response patterns were analyzed in both analyses, the minimum marginal reproducibilities are identical. Given the higher reproducibility for the ordering-theoretic results, the percentage improvement figure is higher for the nonlinear hierarchy than for the linear hierarchy. Finally, the hierarchy generated by the ordering analysis evidenced a coefficient of scalability considerably larger than the hierarchy from the scalogram analysis.

### Conclusions

The results of the study support the use of ordering-theoretic analysis in the evaluation of attitude scales. While nonlinear item hierarchies may not always match the conceptual simplicity of Gutt-

TABLE 4  
*Comparison of Scalogram and Ordering-Theoretic  
Results on Hierarchy Validation Statistics*

Statistic	Scalogram Analysis	Ordering Analysis
Reproducibility	.84	.95
Minimal Marginal Reproducibility	.72	.72
Percentage Improvement	.12	.23
Coefficient of Scalability	.43	.82



man scales, it is likely that nonlinear hierarchies provide a more accurate representation of the inter-item relationships existing in various attitudinal domains. The depiction of a branch hierarchy of specific logical relationships among attitudinal items, rather than the simple linear sequence afforded by scalogram analysis, may provide the researcher with more insight into the dynamics underlying the attitude.

Whenever logical relationships between test items or tasks are of interest, ordering theory can be used. Ordering theory can reveal nonlinear lines of implication among items or tasks and in so doing, serve as a basis for hypothesizing lines of causation to be tested in experimental settings.

## REFERENCES

- Airasian, P. W. The role of evaluation in mastery learning. in J. Block (Ed.) *Mastery learning: Theory and Practice*. New York: Holt, Rinehart and Winston, 1971, 77-88. (a)
- Airasian, P. W. A method for validating sequential instructional hierarchies. *Educational Technology*, 1971, 11, 54-56. (b)
- Airasian, P. W. A study of behaviorally dependent, classroom taught task hierarchies. *Educational Technology Research Report Series*, No. 3, 1971 (c)
- Airasian, P. W. and Bart, W. M. Ordering theory: A new and useful measurement model. *Educational Technology*, 1973, 13, 56-60.
- Airasian, P. W., Bart, W. M., and Greaney, B. J. An ordering theoretic analysis of a propositional logic game. Paper read at the annual American Educational Research Association meeting, New Orleans, La., February 1973.
- Anderson, R. E. A computer program for guttman scaling with the goodenough technique. *Behavioral Science*, 1966, 2, 235.
- Bart, W. M. and Airasian, P. W. The generation of item hierarchies by an ordering-theoretic method and a piagetian example. Paper read at the annual American Educational Research Association meeting, Chicago, Ill., April, 1972.
- Bart, W. M. and Krus, D. J. An ordering-theoretic method to determine hierarchies among items. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1973, 33, 291-300.
- Cox, R. C. and Graham, G. T. The development of sequentially scaled achievement test. *Journal of Educational Measurement*, 1966, 3, 147-150.
- Gagné, R. M. The acquisition of knowledge. *Psychological Review*, 1962, 69, 355-365.
- Gagne, R. M. Curriculum research and the promotion of learning. *Perspectives of Curriculum Evaluation*. Chicago: Rand McNally, 1967, 19-38.
- Gagne, R. M. Learning hierarchies. *Educational Psychologist*, 1968, 6, (1).

- Gagne, R. M. and Paradise, N. E. Abilities and learning sets in knowledge acquisition. *Psychological Monographs*, 1961, 75 (whole number 518).
- Goodstein, R. L. *Boolean algebra*. London: Pergamon Press, 1963.
- Guttman, L. A. A basis for scaling qualitative data. *American Sociological Review*, 1944, 9, 139-150.
- Guttman, L. A. A basis for scalogram analysis. In Stuffer, et al., *Studies in social psychology in World War II, Volume 4, Measurement and prediction*. Princeton: University Press, 1950.
- Kerlinger, F. N. The first and second order factor structures of attitudes toward education. *American Educational Research Journal*, 1967, 3, 191-205.
- Krus, D. J. and Bart, W. M. An ordering-theoretic method of multi-dimensional scaling of items. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1974, 34, 525-535.
- Lazarsfeld, P. F. Latent structure analyses. In Koch (Ed.), *Psychology: a study of a science*, Vol. 3. New York: McGraw Hill Book Co., 1959.
- Lele, K. and Bart, W. M. Preliminary item analysis by ordering theory. University of Minnesota Research and Development Center in Education of Handicapped Children, Minneapolis, Minnesota, 1971.
- Lingoes, J. C. Multiple scalogram analysis: a set theoretic model for analyzing dichotomous items. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1963, 23, 501-524.
- Torgerson, W. S. *Theory and methods of scaling*. New York: Wiley, 1958.
- Walbesser, H. H. A hierarchically based test battery for assessing scientific inquiry. Paper read at the annual American Educational Research Association meeting, Los Angeles, California, February 1969.



## THE DIFFERENTIAL FORMATION OF RESPONSE SETS BY SPECIFIC DETERMINERS<sup>1</sup>

PHILLIP D. JONES<sup>2</sup> AND GARY G. KAUFMAN  
Internal Revenue Service

This investigation was conducted to determine if both position and alternative length specific determiners cause a differential formation of response sets on tests in high and low scoring groups. A 46 item vocabulary test with 10 alternate forms varied by type and frequency of specific determiners was administered to 1000 undergraduate Psychology students. Results indicated that as the frequency of specific determiners increased, they formed increasingly strong but differential guessing response sets in high and low scoring groups; however, the magnitude of the effect was much stronger for position specific determiners. Results were interpreted in terms of differing frequencies of appearance in existing tests and the actual nature of the responses an examinee makes to multiple-choice items.

IN the mid-1920's, educators perceived the value of using objective examinations as a method for testing the achievement or aptitude of individuals and published a variety of articles and books which discuss specific recommendations for objective item and test construction (Weideman, 1926; Orleans and Sealy, 1928; Odell, 1928). Included in the ensuing discussion of the technical considerations that must be dealt with in the writing of objective test items was more caution to avoid providing clues or "specific determiners" which raised above the chance level the probability of success on an item about which the examinee has no knowledge (Lang, 1930; Hawks, Linquist, and Mann, 1936; Remmers and Gage, 1955; Travers, 1955; Garrett, 1964; and Stanley, 1964).

<sup>1</sup> The authors wish to thank Professor E. E. Cureton for his invaluable guidance and critical review during the preparation of this study.

<sup>2</sup> Requests for reprints should be sent to Phillip Jones, Internal Revenue Service, Room 5501, 550 Main St., Cincinnati, Ohio 45202.

During the late 1920's, other investigators began to examine the existence of response tendencies associated with examinees rather than specific item construction. These response tendencies, now called response sets, seem to cause "a person consistently to give different responses to items than he would when the same content is presented in a different form" (Cronbach, 1946).

During the next 30 years, response sets and specific determiners were treated as rather independent entities. Moreover, test writers seemed to agree with Cronbach (1946) that multiple-choice objective examinations were free from response sets. However, in 1962, Wevrick performed the first empirical investigation of the possibility of an interaction between the two entities.

Wevrick's study was designed "to determine if conditions could be contrived such that positional response biases would be demonstrated in a highly structured multiple-choice test." Wevrick's hypotheses were: (a) for each item, if the correct alternative occupies the biased position, then the proportion of correct responses made to that item will increase; and, (b) if the position of the correct alternative is randomly distributed across items, a (position) response set will not influence the total score distribution.

Wevrick successfully established a response set for the keyed position by utilizing a test whose arrangement led to a single position being the keyed answer nearly all of the time for the first (easiest) test items and thereafter decreasing in frequency of appearance. However, no such response set was established for tests involving randomly keyed alternatives.

A two-part study reported by Chase (1964) demonstrated the existence of a *sui-generis* response set to choose markedly longer alternatives in a multiple-choice test on a topic about which a sample of college students had no knowledge. A second part of the study further demonstrated that with a carefully constructed test, this preexisting response set could be removed.

To summarize, Wevrick's and Chase's studies seem to imply that there is a causal relationship between specific determiners and the formation of response sets. If this relationship does in fact exist, then any test which contains specific determiners may have its reliability spuriously inflated by correlated, but irrelevant, consistencies and its validity lowered since those consistencies may be uncorrelated with those abilities the test is trying to assess. It would also seem logical that the formation of these sets would be a direct function of the amount of exposure an examinee had to the operation of the specific determiner and that more knowledgeable students will have more exposure and thus be differentially aided by more educated guessing.



The purpose of this study, therefore, is to determine if in fact there is a differential formation of response sets in groups having high and low test scores through the action of specific determiners and to determine if this formation of response sets occurs with two types of specific determiners.

### *Methods*

#### *Construction—Position Specific Determiner*

The general construction of all forms examining the differential formation of response sets in high and low performance groups by position specific determiners was similar. Each test consisted of 46 four alternative multiple-choice vocabulary items in which the examinee was to choose the alternative which was most closely synonymous to the stem word. Nine of the items on each test were "pure guess" items in which the stem word was selected from a list of words given in the *Teacher's Word Book of 30,000 Words* (Thorndike and Lorge, 1944) as appearing four times in 18,000,000 occurrences and none of the alternatives were synonyms or antonyms of the stem word.

The items in each test were arranged so that each of the first four items was keyed in a different position and the next two items were "pure guess" or probe items. This arrangement allowed the research to determine, by examining the probe items, whether or not an initial position response set was in existence. The remaining 40 items, including seven probe (or "pure guess") items spaced evenly throughout, were keyed according to the following schedule:

- Form 0—twenty-five percent of the items were keyed in each position
- Form 4—forty percent of the items were keyed in position two, and the rest of the keyed responses were equally distributed among the other three alternatives
- Form 5—fifty percent of the items were keyed in position two, and the rest of the keyed responses were equally distributed among the other three alternatives
- Form 6—sixty percent of the items were keyed in position two, and the rest of the keyed responses were equally distributed among the other three alternatives
- Form 7—seventy percent of the items were keyed in position two, and the rest of the keyed responses were equally distributed among the other three alternatives

All items with a keyed response were chosen from a list of one hundred items provided by Ruff (1967) who developed response dis-

crimination indices for these items in a sample of junior and senior level high school students. Stem words and alternatives with the highest response discrimination indices (alternative-total score correlations ranging from .40 to .78 for the correct alternatives and from  $-.11$  to  $-.67$  for the incorrect alternatives) were selected to be included in this study.

Each form of the test consisted of the same items, with alternatives arranged in an order that would provide the desired percentage occurrence of the position specific determiner.

### *Administration—Position Specific Determiner*

One hundred copies of each form were administered without time limit to separate sections of an Introductory Psychology class at the University of Tennessee. There were no oral instructions. Instructors in each of these sections had asked one hundred students to participate in a graduate research study. Written instructions on the first page of each form were as follows:

The following test attempts to measure intelligence by responses indicating knowledge of the meaning of abstract words. There are a few easy items which nearly everyone will get right and a few very hard items which almost no one will answer correctly. For our research purposes, we need to have an answer for every item on the test, so no matter how unfamiliar a word may be, please record the best response you can, even if it is a pure guess. It is critical to our research that you work straight through the test and not skip around in answering the questions. Please place all your answers on the answer sheet. Do not write on the test itself. ON YOUR ANSWER SHEET PLACE THE NUMBER OF THE WORD WHICH HAS THE MEANING CLOSEST TO THE FIRST (CAPITALIZED) WORD FOR EACH ITEM.

### *Scoring—Position Specific Determiner*

The scoring on all forms was as follows:

1. The number of correct, keyed items was determined for each of the 100 subjects.
2. The top 25 scores and bottom 25 scores were selected as constituting the High and Low Groups, respectively. In the case of ties for the 25th score in each group, a random selection was made among all those tied scores.
3. The alternative selected for each of the "pure guess" items was tabulated separately for all members in each of the two groups.

### *Construction—Longest Alternative Specific Determiner*

The general construction of all tests examining the differential formation of response sets in High and Low performance groups by alternative length specific determiners was similar to that of the forms previously discussed.

Each test consisted of 46 four alternative multiple-choice vocabulary items in which the examinee was to choose the alternative which provided the best definition of the stem word. Each of these items was formed by simply providing the definitions of the synonyms used on the forms for testing position specific determiners.

Each form of test consisted of the same items, with the longest alternative being the keyed response according to the following schedule:

Form 00—the longest alternative was keyed 25% of the time.

Form 8—the longest alternative was keyed 40% of the time.

Form 9—the longest alternative was keyed 50% of the time.

Form 10—the longest alternative was keyed 60% of the time.

Form 11—the longest alternative was keyed 70% of the time.

Care was taken to ensure that any position was keyed only on twenty-five percent of the items and that each "longest" alternative was at least twice as long as any other alternative (based on actual word count).

### *Administration—Longest Alternative Specific Determiner*

Procedures for administering these forms were identical to those for administering forms examining position specific determiners with the exception that the final sentence of instructions was changed to read, "ON YOUR ANSWER SHEET PLACE THE NUMBER OF THE DEFINITION WHICH MOST CLOSELY DEFINES THE CAPITALIZED WORD FOR EACH ITEM."

### *Scoring—Longest Alternative Specific Determiner*

Procedures for selecting the 25 members of the High Group and Low Group were identical to those for the position specific determiner test forms. After these groups were formed, tabulation was made whether or not the longest alternative was selected for each "pure guess" item for all individuals in both groups.

### *Results*

Data (Forms 0,4,5,6, and 7) examining the prior existence of a response set to guess position number two were analyzed using a *t*-test

for the mean number of times position two was selected on the first two probe items against the mean value expected by chance. The computed  $t$ -statistic was 0.50, a value not significantly different from chance.

Data examining the formation of response sets by position specific determiners were initially analyzed using a one-factor analysis of variance for probe items three through nine, in which the main effect was the percentage occurrence of items keyed in position number two on each form of the test. Separate analyses were conducted for groups designated as "High" and "Low" based on their number of correct responses on non-probe items. Results of the separate analyses demonstrated a significant main effect in both the High Group ( $F = 5.62, p < .001$ ) and Low Group ( $F = 2.53, p < .05$ ).

Further analyses were completed by computing  $t$ -statistics for all forms within each group, testing the difference between the mean number of times position two was selected by individuals within each group against the mean number of times that alternative would be expected to be selected by chance.

As demonstrated by Table 1, all means were ordered in the expected direction in both groups.

A sign test computed on this ordering was significant at  $p = .031$  for each group. Also, neither of the two control groups (Form 0, High Group and Low Group) selected position number two significantly more often than expected by chance; whereas, with one exception, position two was selected significantly more often than chance would predict in all other groups. Illustration of these data are provided in Figure 1.

Additional  $t$ -statistics were computed that tested the differences

TABLE 1  
*t*-Tests for Position Two against Theoretical Mean for High and Low Groups

High Groups	Observed Mean	Theoretical Mean	$t$	$p$
Form 0	1.76	1.75	.01	NS
Form 4	2.64	1.75	2.81	.005
Form 5	2.64	1.75	3.02	.005
Form 6	3.04	1.75	6.34	.001
Form 7	3.48	1.75	8.04	.001
Low Groups				
Form 0	1.60	1.75	-.71	NS
Form 4	2.16	1.75	2.28	.05
Form 5	2.20	1.75	1.62	NS
Form 6	2.44	1.75	2.90	.005
Form 7	2.64	1.75	3.09	.005

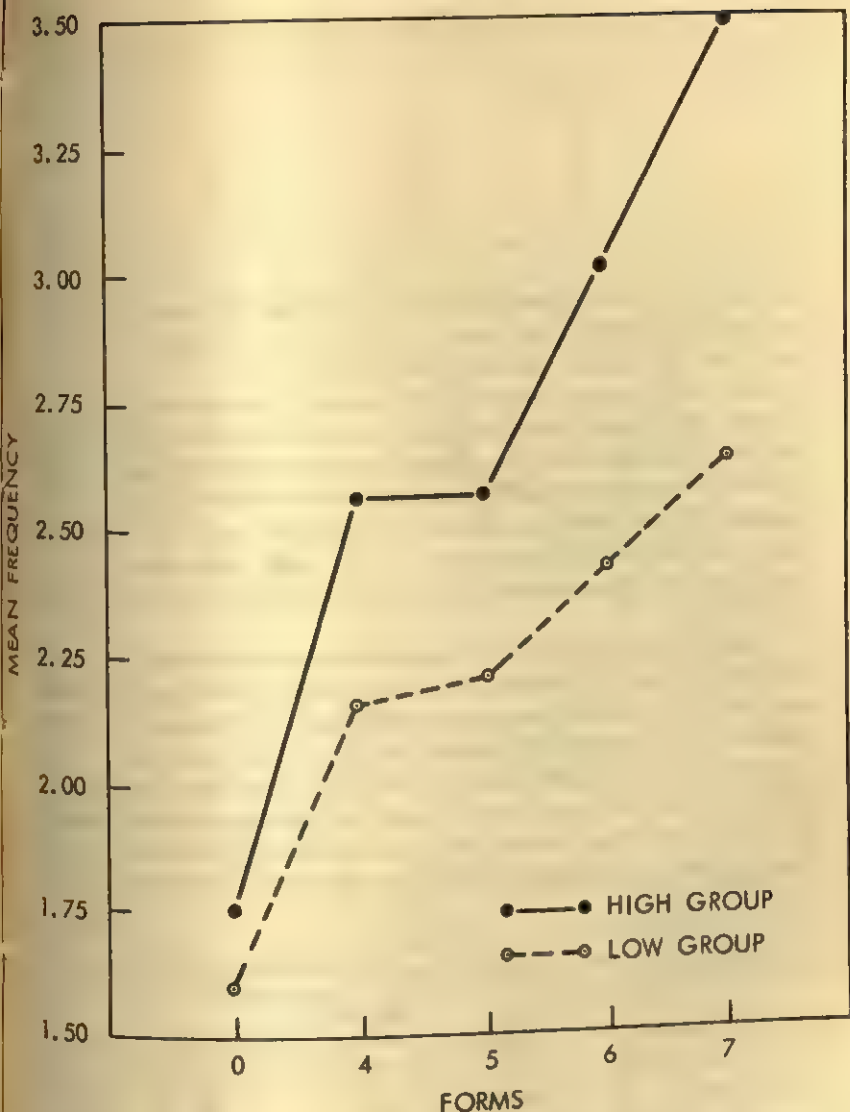


Figure 1. Mean frequencies for selection of position two in High and Low groups.

between the mean number of times position number two was selected in the High and Low Groups on each form. As shown in Table 2, significant differences existed between these groups on Forms 6 and 7.

Data (Forms 00, 8, 9, 10, and 11) examining the prior existence of a response set to guess the longest alternative were analyzed using a *t*-test comparing the mean number of times that alternative was selected



TABLE 2  
*t*-Tests for Position Mean Differences between High and Low Groups

Form	<i>t</i>	<i>p</i>
0	.53	NS
4	1.26	NS
5	1.10	NS
6	1.71	.05
7	2.21	.025

on the first two probe items to the mean number of times it would be expected to be selected by chance. The computed *t*-statistic was 0.50, a value not significantly different from chance.

Data examining the formation of response sets by overkeying the longest alternative were analyzed using a one-factor analysis of variance for probe items three through nine in the same manner as that for the formation of response sets by position specific determiners. Results of these analyses indicate a significant main effect between forms for the High Groups ( $F = 3.31, p < .05$ ), while there was no significant difference between forms for the Low Groups ( $F = 0.45, p = NS$ ).

Computations of *t*-statistics testing the difference between the mean number of times the longest alternative was selected by individuals with each group against the number of times that alternative would be expected to be selected by chance are presented in Table 3.

As shown by this table, only in one of the ten groups is the longest alternative selected significantly more often than expected by chance. However, all means within the High Group are ordered in the predicted direction. Illustration of these data is provided in Figure 2.

TABLE 3  
*t*-Tests for Longest Alternative against Theoretical Mean for High and Low Groups

High Groups	Observed Mean	Theoretical Mean	<i>t</i>	<i>p</i>
Form 00	1.44	1.75	-2.17	.05
Form 8	1.64	1.75	.53	NS
Form 9	1.88	1.75	.49	NS
Form 10	2.08	1.75	1.13	NS
Form 11	2.64	1.75	2.65	.025
Low Groups				
Form 00	1.43	1.75	-1.32	NS
Form 8	1.68	1.75	-.38	NS
Form 9	1.52	1.75	-1.14	NS
Form 10	1.80	1.75	.24	NS
Form 11	1.76	1.75	.05	NS

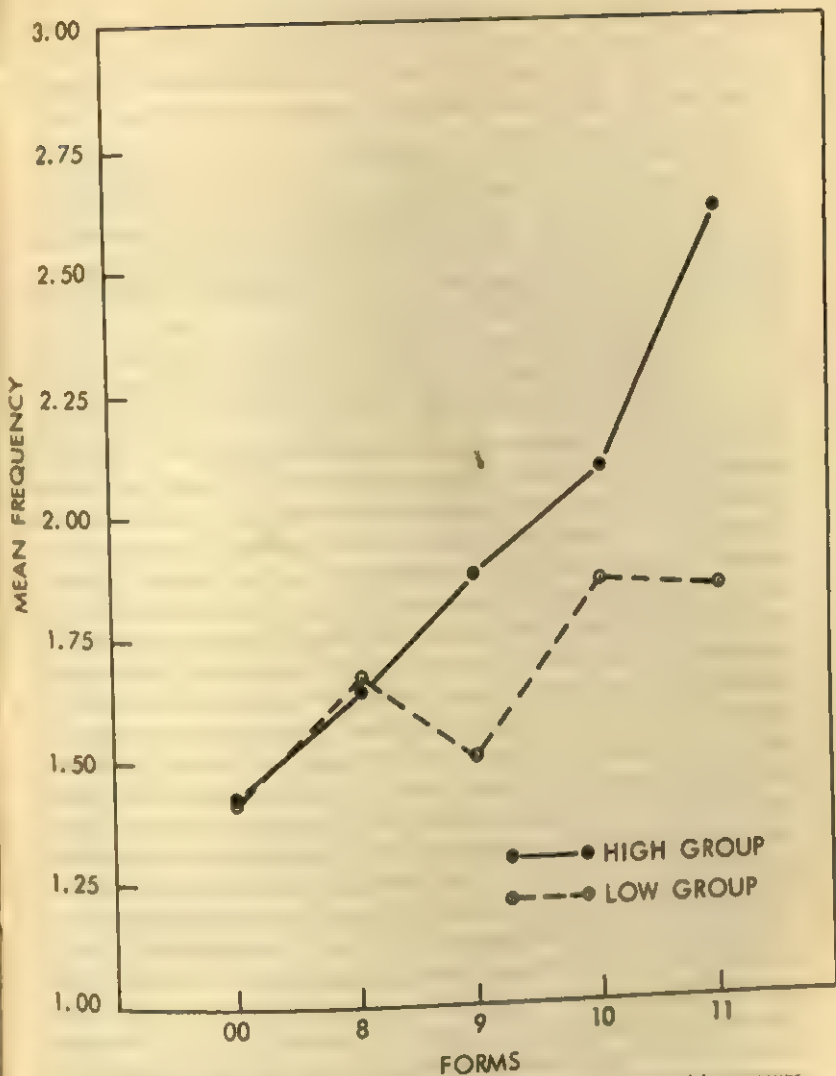


Figure 2. Mean frequencies for selection of longest alternative in High and Low groups.

Since the longest alternative was selected significantly more often than dictated by chance in Form 11 only, an additional *t*-statistic testing the difference between the mean number of times the longest alternative was selected in the High and Low Groups was computed only for that form. The computed *t*-statistic was 2.61, significant beyond the .01 level.

*Discussion*

The results of the analysis of data for both position and alternative length specific determiners have led the authors to four conclusions:

1. Specific determiners can cause the formation of response sets.
2. The strength of a response set for a given test is positively related to the number of correct answers an individual makes on that test. Thus, differential response set formation occurs.
3. The strength of a response set varies with the extent to which the specific determiner occurs within a test.
4. The strength of a response set appears to vary according to the type of specific determiner, e.g., the position response set was more easily formed than the alternative length response set.

Why have these phenomena occurred? Wevrick's (1962) discussion implies that the formation of response sets by specific determiners is an instrumental conditioning phenomenon with partial reinforcement operating to sustain the response set. The reasoning behind this instrumental conditioning approach seems to be that the test item alternatives act as a stimulus condition, marking an alternative serves as the operant response, and knowing that one has marked the correct alternative is the reinforcement. Through the operation of specific determiners, the response of marking a specific alternative is reinforced by knowledge that the alternative is correct, and thus the occurrence of marking that alternative is increased. However, the major problem with this interpretation is that, in the test situation, the reinforcement precedes the conditioned response rather than following it. That is, the examinee's reason for marking a particular response is that he already believes it is the correct one. Therefore, the situation is in clear violation of the contingency principle of instrumental conditioning.

The investigators feel that a classical conditioning analysis provides a clearer explanation of why the phenomenon occurs. In this analysis, the conditioned stimulus, position two, occurs prior to the onset of the unconditioned stimulus, recognizing the correct alternative. The pairing of position two with recognition of the correct response for a number of trials appears to cause that position to elicit the conditioned response of marking it given the absence of a conflicting stronger response (recognizing the correct alternative in some position other than two).

This classical conditioning analysis would seem to provide an explanatory scheme for (a) the formation of response sets; (b) the positive relation between the strength of response sets and the number of correct answers (the higher scorers have more conditioning trials

and fewer extinction trials); and (c) the positive relationship between the strength of the response set and the extent to which the specific determiner occurs (more learning trials).

However, the analysis does not explain why the position specific determiners established a stronger response set than the alternative length specific determiner.

There seems to be two plausible reasons why position specific determiners form strong guessing response sets while response length specific determiners do not. The first is that since these data have in fact demonstrated that the formation of guessing response sets by specific determiners is a function of the amount of exposure to that specific determiner, and since analyses by Jones (1972) demonstrated that position specific determiners occur in both published aptitude and achievement tests and teacher-made tests far more frequently than do response length specific determiners, one may logically conclude that examinees have been more exposed to this type of specific determiner and are more prone to perceive it and make their guesses accordingly.

A second explanation for these results is based on the difference between the alternative selected acting as a stimulus to or a response of the individual examinee. Position specific determiners act not only as stimuli for the subject's response, but more importantly, act as a response themselves. That is, examinees actually make the response of marking position two, for example, on their answer sheets. On the other hand, the longest alternatives function only as stimuli for a response. That is, even though the examinee selects the longest alternative as the correct answer, his response is still the marking of the position that alternative occupies on his answer sheet. Thus, it might very well be speculated that the key element in the formation of response sets by specific determiners is the response the subject makes, not the stimulus for that response.

This hypothesis can be tested in future research by simply having the subject write down on the answer sheet the alternative he has chosen. Thus, when probing for the formation of guessing response sets by response length specific determiners, if the subject writes down the longest alternative, it serves not only as a stimulus, but as his response. If the hypothesis holds true, one would predict the strong differential formation of guessing response sets to occur as a result of the longest alternative specific determiner just as for the position specific determiner.

The practical significance of the findings of this study depends upon the extent to which specific determiners occur in actual tests, both teacher-made and professionally constructed published examinations.

Two studies cast some light on this matter. Metfessel and Sax (1958)

examined 30 multiple-choice and true-false standardized published achievement and aptitude tests and tabulated the position of the keyed responses. They obtained Chi-squares significant at the .01 level of probability or better for 42% of the tests they examined and concluded that there seemed to be a general trend in multiple-choice tests for the keyed answer to be in the middle of the distribution, and in the true-false tests, test makers tended to write more true questions than false.

More recently, Jones (1972) examined 105 teacher-made college course tests and 49 published aptitude and achievement tests for the existence of position or alternative length specific determiners. A Chi-square analysis applied to the data gathered in this study demonstrated that 37 of the teacher-made tests and five of the published tests rejected the hypothesis that "no position is keyed more often than dictated by chance" at least at the .05 level of probability. It was stated, however, that it was suspected that for some of the published tests, only the relatively small number of items in a sub-test prevented the Chi-square analysis from being significant. Support for the supposition was provided by the fact that ten of the sub-tests had one alternative that was never keyed and yet had a computed Chi-square that did not differ significantly from chance.

This study further demonstrated that six of the teacher-made tests and none of the published tests seemed to involve the significant overkeying of a "clearly longer" alternative.

Thus, it seems likely that specific determiners do exist in real testing situations and they often exist in sufficient concentration to cause the formation of a guessing response set.

The practical significance of this study is apparent. The failure to eliminate such specific determiners will result in spurious measurement of reliability and validity due to the inclusion of a measured correlated error and will in addition create a bias in favor of more knowledgeable students.

## REFERENCES

- Chase, C. I. Relative length of options and response set in multiple-choice items. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1964, 24, 861-866.
- Cronbach, L. J. Response sets and test validity. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1946, 6, 475-494.
- Garrett, H. E. *Testing for teachers*. New York: American Book Company, 1964.
- Hawks, H. E., Lindquist, E. F., and Mann, C. R. *The construction and use of achievement examinations*. Boston: Houghton Mifflin Company, 1936.



- Jones, P. D. *The differential formation of response sets by specific determiners*. (Doctoral dissertation, University of Tennessee) Knoxville, Tennessee, 1972.
- Lang, A. R. *Modern methods in written examinations*. Boston: Houghton Mifflin Company, 1930.
- Metfessel, N. S. and Sax, G. Systematic biases in the keying of correct responses on certain standardized tests. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1958, 18, 787-790.
- Odell, C. W. *Traditional examinations and new type tests*. New York: The Century Company, 1928.
- Orleans, J. S. and Sealy, G. A. *Objective tests*. Chicago: World Book Company, 1928.
- Remmers, H. H. and Gage, H. L. *Educational measurement and evaluation*. New York: Harper and Brothers, 1955.
- Ruff, R. D. Reliability and the systematic elimination of response alternatives. Unpublished Masters Thesis; University of Tennessee, 1967.
- Stanley, J. C. *Measurement in today's schools*. Englewood Cliffs: Prentice-Hall, 1964.
- Thorndike, R. L. and Lorge, I. *Teachers world book of 30,000 words*. New York: Teacher's College Press: Columbia University, 1944.
- Travers, R. M. W. *Educational measurement*. New York: Macmillan Company, 1955.
- Weideman, C. C. How to construct the true-false examination. *Teacher's College Contributions to Education*, 1926, No. 225.
- Wevrick, L. Response set in a multiple-choice test. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1962, 22, 533-538.



## STRUCTURE OF ACADEMIC ATTITUDES AND STUDY HABITS

S. B. KHAN

The Ontario Institute for Studies in Education

DENNIS M. ROBERTS

The Pennsylvania State University

SSHA was administered to a total of 846 students. Their responses were analyzed to test the a priori classification of items into DA, WM, TA, and EA scales and to test the hierarchical structure of the scales. Transformations of the initial factor matrix to varimax and congruence to a hypothesized structure supported the classification of items into DA, WM, and TA scales but not the EA scale. The higher-order factoring of the initial factor-intercorrelations revealed a two-stage hierarchical structure. Implications for students' counseling are discussed.

THE evaluation of students' attitudes and study habits has, in part, been prompted by the inability of aptitude variables to account for a major portion of variance in school learning and achievement. Brown and Holtzman (1953) developed the *Survey of Study Habits and Attitudes* (SSHA) for measuring relevant attitudes and study habits and recommended its use for diagnosis, counselling, and prediction. One of the major criticisms of the SSHA was that it yielded a single score which did not contain much information about the strengths and weaknesses of an individual in specific areas. To alleviate this criticism, Brown and Holtzman (1965) revised the instrument and classified the items on the basis of judgments by 15 experts into four a priori scales; namely, Delay Avoidance (DA), Work Methods (WM), Teacher Approval (TA) and Education Acceptance (EA). A Study Habits (SH) score is then obtained by combining scores on DA and WM scales and a Study Attitudes (SA) score is obtained by combining

scores on TA and EA scales. Finally, a total score of Study Orientation (SO) is obtained by further combining the SH and SA scores. The combination of scales was justified on the basis of intercorrelations among the four basic scales for a sample of 3,054 college freshmen (Brown and Holtzman, 1966). For instance, scores on DA and WM scales correlated .70 with each other and scores on TA and EA scales correlated .69 with each other. Other correlations were: DA vs TA = .49, DA vs EA = .65, WM vs TA = .53, WM vs EA = .62.

Although not explicitly stated, the authors seem to have forwarded the notion of a hierarchy of academic attitudes and study habits. This hierarchy would have the SO factor as the top of the pyramid, the SH and SA factors as second tier, and the DA, WM, TA and EA scales as the branches off to the third tier. It was the purpose of the present study to empirically assess the validity of the a priori scales and to test the hypothesis of a hierarchical structure of attitudes and study habits.

### *Method*

The sample for the study came from senior classes in four high schools ( $N = 243$ ) and from freshman classes in two universities ( $N = 603$ ) in Ontario. Each respondent was asked to indicate whether a statement on the SSHA was rarely (0 to 15% of the time), sometimes (16 to 35% of the time), frequently (36 to 65% of the time), generally (66 to 85% of the time), and almost always (86 to 100% of the time) true for him. The five options of rarely to almost always were assigned numerical values from 1 to 5 respectively. Data for the school and university samples were combined because the SSHA, Form C is intended for both high school seniors and college freshmen.

Pearson product-moment correlations were obtained among the items by assuming that the response scale was continuous and that responses to each item were normally distributed in the population from which the the samples were drawn. To determine whether covariation among responses can be explained by the four a priori scales, the inter-item correlation matrix was analyzed by the method of principal components. Four components associated with the first four eigenvalues were transformed to simple structure by the normalized varimax procedure and the resulting transformed solution was psychologically interpreted by noting the proportion of items which had loadings (arbitrarily selected as .35) higher than the critical value on the appropriate factors.

The same question was examined in another way. A factor matrix was constructed by placing 1's, -1's and 0's as elements of the matrix according to the classification of items in each scale and whether an

item was positively or negatively worded. The *observed* factor matrix was transformed orthogonally to similarity to the *constructed* factor matrix by a procedure from Cliff (1966). Coefficients of congruence ( $\phi pq$ ) between the corresponding factors of the constructed and the observed factor matrices were obtained in order to determine the extent to which a priori classification of items into scales was confirmed by the data.

To test the hypothesis of hierarchical structure, higher-order factors were derived by factoring the correlations among the lower-order factors. In order to facilitate psychological interpretation, the final matrix consisting of the loadings of the variables on the initial and higher-order factors was transformed to an orthogonal position by using the procedure described by Schmid and Leiman (1957).

### Results

The results of the varimax and Schmid-Leiman transformations of the initial factors of SSHA appear in Table 1. The factors under Arabic numbers are the varimax-transformed while those under Roman numerals resulted from the Schmid-Leiman transformation of the second-and first-order factors. The factors have been arranged such that they correspond to the a priori scales. The description of the varimax results follows immediately while that of the Schmid-Leiman transformation is included in the discussion of the hypothesis regarding the hierarchical structure of attitudes and study habits.

According to how the SSHA is constructed in terms of order of items and the corresponding scales they refer to, it is expected that the first and every fourth item thereafter should have higher loadings (.35 or higher) on the first factor (DA), the second and every fourth item should have higher loadings on the second factor (WM), the third and every fourth item should have higher loadings on the third factor (EA). An examination of the loadings on the first factor indicates that 16 out of 25 (64%) items satisfy the above criterion. There are 17 out of 25 (68%) items which load .35 or higher on the second factor. For the third factor, 20 out of 25 (80%) items meet the criterion of having .35 or higher loadings. The proportion of appropriate loadings on the fourth factor is negligible (12%) compared to the other three factors.

Interpretation of a factor is facilitated if the proportion of inappropriate loadings on the factor is relatively small. Theoretically, there could have been as many as 75 inappropriate loadings yields 6.66% on each factor. A count of the inappropriate loadings yields 6.66% on the first factor, 1.33% on the second factor, 10.66% on the third factor, 26.66% on the fourth factor.



TABLE I  
*Varimax and Schmid-Leiman Transformation of SSHA Factors<sup>a</sup>*

ITEMS	I <sup>b</sup>	I <sup>c</sup>	II	2	Factors III	3	IV	4	V
1	46	45							
2									
3	40					63	50	50	49
4	40								
5									
6				67	66			43	38
7	-43					-44		36	
8									
9	45	52	35						
10				41	41			-41	-38
11	41					54	38		
12	-41	-41						52	42
13	-36	-37						49	40
14								62	59
15						35			
16	48								
17								47	42
18				43	36				
19	43					70	55		
20									
21								37	
22									
23	48					73	56		
24						48	36		
25	39								
26				55	47			-35	
27	42					50			
28	56	41							
29								39	
30				56	57				
31				50	46				
32	48	35							
33	35	47						-38	
34				53	48				
35	45					52	35		
36									
37	53	67	47						
38				58	51				
39	42					70	56		
40	53	36				42			
41	50	49							
42	35			35					
43	41					59	44		
44	38					36			
45	44	55	37						
46	35			55	46				
47	35					45			
48	35					49	37		
49	58	59	38						
50				65	59				
51						-44			
52	-38								
53	50	65	44						
54									

TABLE 1 (Continued)

55	43					50	35		
56									
57	56	60	38					-35	
58								44	41
59	39					56	41		
60	55	46				41			
61	37	52	37						
62				56	52				
63	35					45			
64	40					45			
65								44	37
66	37			60					
67									
68									
69	43	45							
70								50	46
71	38					57	43		
72	-41							59	51
73	-37	-35						54	45
74				46	37				
75									
76									
77									
78								53	49
79									
80									
81	54	63	42						
82				46	46				
83	36					50	38		
84									
85	-39	-45							40
86				40	38				
87									
88	37					49	36		
89	-47	-53						51	39
90								43	39
91	35					46			
92									
93									
94	38			46					
95	37					45			
96	49	47							
97	-38							57	50
98				48	39				
99						-38			
100	38					43			
App.						20			3
%		16		17		80			12
Inapp.		64		68		8			20
%		5		1		10.66		26.66	
$\phi_{pq}$		6.66		1.33		.66		.28	
		.54		.71					

\* Decimal points omitted and loading less than .35 not reported.

<sup>b</sup> Transformed varimax factors.

<sup>c</sup> Transformed Schmid-Leiman factors.

The results of the analyses of transformation of the observed factor matrix to the constructed matrix yielding coefficients of congruence of .54, .71, .66, and .28 for the four factors. The results of the simple counting of appropriate and inappropriate factor loadings are substantiated by the magnitude of the coefficients of congruence. The sampling distribution of the coefficient of congruence is not known which makes it difficult to determine whether the above values are statistically significant. Evans (1970) suggests that coefficients of .90 or higher indicate "good" correspondence, coefficients from .80 to .90 show "fair" correspondence, coefficients from .70 to .80 indicate "poor" correspondence between a pair of factors. According to this rule of thumb, there seems to be poor correspondence at best between the hypothesized and observed factors, however, this rule may be too stringent to apply to an item factor matrix because correlations among single items may not be expected to be as high as correlations among tests or subtests containing a reasonable number of items. Keeping this in mind, the results of similarity analysis may be interpreted to mean that the data generally support the a priori classification of items for three out of the four scales.

The results of the factoring of the four first-order factors do not confirm the type of hierarchical structure proposed for study habits and academic attitudes. The group factors of SH and SA did not emerge; instead, a general factor (Study Orientation) resulted from the analysis of inter-relationships among the four factors. The present results support a two-stage rather than a three stage hierarchy of academic attitudes and study habits as measured by SSHA.

The general factor is mainly composed of items from the DA and TA scales. The WM and the fourth scale seem to be specific and have not as much in common with the general factor as the other two scales. This observation is reflected in the intercorrelations and the factor loadings of the first-order factors on the general factor presented in Table 2. For instance, the general factor explains 61% and 53% of the variance in the DA and TA scores respectively compared to 27% and 18% of of the variance that is common to the general factor and WM and the fourth factor respectively.

### *Discussion*

The results tend to suggest that a priori classification of items hold for DA, WM, TA, scales but not for the EA scale. An analysis of the content of items which load on factor 4 reveals a tendency to apply one's self seriously to systematic studying and to doing assignments. These items were judged to belong in the DA and WM categories. It

TABLE 2  
*Intercorrelations among First-Order Factors and Their  
 Loadings on the Second-Order Factor<sup>a</sup>*

First-Order Factors	1	2	3	4	I <sup>b</sup>
1 Delay Avoidance		41	52	-41	78
2 Work Methods			44	-10	52
3 Teacher Approval				-31	73
4 Academic Diligence <sup>c</sup>					-42

<sup>a</sup> Decimal points omitted.

<sup>b</sup> Second-order factor.

<sup>c</sup> Renamed.

seems that these items tap motivational characteristics rather than modes of studying and factor 4 may be interpreted as a measure of "academic diligence."

Twelve of the items, which were supposed to have loaded on an EA factor, did not load on any of the four factors. In an earlier study (Khan, 1969), almost similar items were hypothesized to measure an "attitude toward education" factor for Junior High School students. Such a factor did not emerge in separate analyses for both males and females. Most of the remaining items classified in the EA scale have appeared on the TA factor with a few items loading on the DA factor. A scrutiny of these items indicates that there is an implicit or explicit reference to the teacher, and the respondents have interpreted these statements in relation to the teacher as the stimulus object.

The intercorrelations among the subscales reported by Brown and Holtzman (1966) do not seem to justify a three-stage hierarchy of school-related attitudes and study habits as measured by SSHA. The correlations between the subscales making second-order SH and SA scales are large enough to indicate the presence of a general factor. The present findings confirm such an expectation.

Brown and Holtzman (1966) have emphasized the value of the four subscales in diagnosis and counselling by an analysis of an individual's responses to statements in each subscale. The results of the present study have not supported the existence of an "Education Acceptance" scale and further work on the items making up this scale may be necessary before it is recommended for use in the evaluations of students' attitudes toward education and counselling based upon these evaluations. The present results also indicated that the "Study Habits" and "Study Attitudes" scales do not follow from the basic subscales. Although a counsellor may be interested in knowing whether a student needs help in the area of attitudes or study skills, it is doubtful if this information is provided by adding scores on the "appropriate" sub-

scales of SSHA. It is suggested that scores on the EA subscale and on the SH and SA scales be interpreted with caution because the validity of these subcomponents of SSHA is questionable.

## REFERENCES

- Brown, W. F. and Holtzman, W. H. *Survey of study habits and attitudes*. New York: Psychological Corporation, 1953.
- Brown, W. F. and Holtzman, W. H. *Survey of study habits and attitudes. Form C*. New York: Psychological Corporation, 1965.
- Brown, W. F. and Holtzman, W. H. *Survey of study habits and attitudes manual. Form C*. New York: Psychological Corporation, 1966.
- Cliff, N. Orthogonal rotation to congruence. *Psychometrika*, 1966, 31, 33-42.
- Evans, G. T. Congruence transformations: Procedures for comparing the results of factor analyses involving the same sets of variables. Toronto: The Ontario Institute for Studies in Education, 1970. (Mimeo).
- Khan, S. B. Affective correlates of academic achievement. *Journal of Educational Psychology*, 1969, 60, 216-221.
- Schmid, J. and Leiman, J. M. The development of hierarchical factor solutions. *Psychometrika*, 1957, 22, 53-61.



## A MEASURE OF RELIABILITY USING QUALITATIVE DATA<sup>1</sup>

MENI KOSLOWSKY AND HOWARD BAILIT

Department of Behavioral Sciences and Community Health  
University of Connecticut Health Center  
Farmington, Connecticut

In many research activities, the data is unordered or qualitative. In such circumstances, inter-rater reliability is usually measured by calculating a percentage of agreement score between judges. The present paper expands on an equation first introduced by Goodman and Kruskal for obtaining a reliability measure of one item. This formula determines inter-rater reliability for a series of items across many subjects. The statistic that results is easily interpreted and in many ways is analogous to the conventional reliability for quantitative data.

In many types of research activities, it is necessary to obtain a reliability measure for qualitative or unordered data. The procedures that are presently available cannot handle such data using the classical reliability measures. Finn's (1970) method assumes interval type data, and Goodman and Kruskal's (1954) formula for handling reliability of unordered data is good for only one item at a time. This paper expands on the latter's formulation and discusses an approach for calculating the reliability of a series of items. In this way, the procedure is analogous to the usual reliability determination for an achievement test or an attitude scale.

### *Method*

The Goodman and Kruskal formula states that a measure of

---

<sup>1</sup> This work was supported, in part, by Contract NIH-72-4207 from the Division of Dental Health, National Institutes of Health, Public Health Service. The authors would like to thank Paula Valluzzo and Paula Atwood for their help in collecting the data.

reliability in the unordered case is:

$$\lambda r = \frac{\sum P_{aa} - \frac{1}{2}(PM. + P.M)}{1 - \frac{1}{2}(PM. + P.M)} \quad (1)$$

where,  $P_{aa}$  represents the proportion of elements along the diagonal of a contingency table, and  $PM.$ ,  $P.M.$  represent the marginal proportions associated with the modal category for rows and columns, respectively.

Thus, if two judges classified 100 people into the following contingency table according to their neurotic symptoms, the conventional reliability formulas would not be applicable (see Table 1).

It is obvious that the three classifications of neuroses do not form an ordered scale. For this type of data,  $\lambda r$  provides useful insight into gauging inter-rater reliability.

Goodman and Kruskal say the formula may be interpreted "as the relative decrease in error probability as we go from the no information situation to the other-method-known situation" (p. 758). In our case, as  $\lambda r$  increases the probability of judge II making the same assignment as judge I increases. This measure yields a statistic with much more information than the usually reported "proportion of agreement."

### *The Case of Several Items*

It is possible to extend the Goodman-Kruskal formula to include a series of items. Thus, in the previous example, two judges may be required to assign a group of individuals to one of three categories based on a series of independent skills or abilities.  $\lambda r$  would then represent an average of individual  $\lambda r$ 's across  $N$  items:

$$\lambda r = \frac{1}{N} \sum_i \left[ \frac{\sum P_{aa} - \frac{1}{2}(PM. + P.M)}{1 - \frac{1}{2}(PM. + P.M)} \right] \quad (2)$$

TABLE 1  
*The Use of the Goodman-Kruskal Reliability Formula*

		Judge I			Total
		Obsessive	Hysterical	Phobic	
Judge II	Obsessive	20	10	0	30
	Hysterical	10	25	5	40
	Phobic	5	5	20	30
	Total	35	40	25	
		$\lambda r = \frac{.65 - \frac{1}{2}(.40 + .40)}{1 - \frac{1}{2}(.40 + .40)} = .42$			

The disadvantage of formula (2) is that  $\lambda'r$  becomes indeterminate when both judges agree and assign all subjects to the same category for one or even several items. However, since we are concerned with similarity in judges' ratings, and not item discriminability, a score of "1" can be assigned in this case. An example will clarify the use of the formula.

In a recently completed investigation of a new procedure to evaluate the quality of dental care that patients receive, a 3-point scale 1, 3, and 9 was devised. The value "1" was assigned by a judge if he felt the dentist had failed the requirements of a certain criterion, 3 was assigned if the dentist had passed the requirements of a certain criterion, and 9 was assigned if the judge could not make a decision. The usual reliability formulae are inappropriate because "9" in this data has no ordinal relationship to 1 and 3. All patients were seen by two judges and their dental treatment was rated on a total of 29 criteria,<sup>2</sup> each involving an independent technical skill.

Table 2 presents results from this investigation. Along the rows are the item numbers and the columns contain, for each item, the proportion of agreement, the modal proportion for Judge I and the modal proportion for Judge II. The calculated  $\lambda'r$  equals .27. This indicates that the probability of an error associated with just guessing the modal class (chance level) for each item has been decreased by 27%. Thus, if one guessed Judge II's response by following Judge I's classification exactly, he would be better off than guessing the modal category for Judge II.

### *Discussion and Implication*

Many of the limitations associated with  $\lambda'r$  are also associated with  $\lambda'r$ . (For a full discussion of these see the Goodman and Kruskal paper). As is evidenced from Table 2, the restriction on the denominator for an individual item does not invalidate the procedure. When the degree of comparability between judges is being assessed, perfect agreement is the optimal result.

The meaningfulness of a descriptive measure such as  $\lambda'r$  is quite apparent. The information obtained from this statistic can be used in the same way that decisions are made after a conventional reliability determination. As  $\lambda'r$  increases and approaches unity, more and more confidence can be placed in the present classification scheme. However, as  $\lambda'r$  decreases, the assertion of similarity between the

<sup>2</sup> The description presented here has been abbreviated in order to present only those aspects relevant to this article.

TABLE 2  
*An Example of  $\lambda'r$  as an Indicator of Reliability of Several Items*

Item	$\Sigma Paa$	$PM.$	$P.M$
1	.846	.923	.923
2	.923	.846	.923
3	.769	.846	.615
4	1.000	1.000	1.000
5	1.000	1.000	1.000
6	1.000	1.000	1.000
7	.923	1.000	.923
8	.923	.846	.923
9	.846	.923	.769
10	.846	1.000	.846
11	1.000	1.000	1.000
12	1.000	1.000	1.000
13	.769	.846	.923
14	1.000	.769	.769
15	.923	.769	.692
16	1.000	1.000	1.000
17	.923	1.000	.923
18	.846	.846	.769
19	.692	.769	.769
20	1.000	1.000	1.000
21	1.000	1.000	1.000
22	1.000	.923	.923
23	.846	.923	.923
24	.923	1.000	.923
25	.923	.769	.846
26	1.000	.769	.769
27	1.000	.538	.461
28	1.000	.769	.769
29	.769	.692	.923

two judges' ratings becomes more and more untenable and puts the use of the classification scheme into question.

## REFERENCES

- Finn, R. H. A note on estimating the reliability of categorical data. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1970, 30, 71-76.
- Goodman, L. A. and Kruskal, W. H. Measures of association for cross classifications. *Journal of American Statistical Association*, 1954, 49, 732-764.

## RELIABLE DIMENSIONS FOR WISC PROFILES<sup>1</sup>

ANTHONY J. CONGER<sup>2</sup>

University of North Carolina

JUDITH COHEN CONGER

Duke University

Measures of multivariate reliability are calculated for profiles of WISC subscales on three age groups. Profile dimensions based on reliability considerations are also established and matched across age groups and with factor analytic dimensions. While all possible differences among individual subscales are quite unreliable (about .51), a reduced set of five uncorrelated dimensions can be found with a more satisfactory reliability (about .87). In unrotated form, the maximally reliable dimension is essentially total IQ and the second maximally reliable dimension closely resembles a verbal-performance contrast. Four of the five rotated dimensions give a good match across age groups and with Verbal Comprehension, Relevancy, Perceptual Organization and Maze-specific factors. Guidelines for the interpretation and use of WISC subscale profiles are provided for both clinical and research uses.

THE use of intelligence tests for other than general intellectual diagnoses has been criticized in recent years (Anastasi, 1968); undaunted, clinicians have continued to use various combinations of intelligence test subscales to estimate differential abilities (Gainer, 1965),

---

<sup>1</sup> This research was supported in part by a PHS Research Grant No. MH-10006 from the National Institute of Mental Health, Public Health Service and in part by a NSF University Science Development Program award No. GU-2059 and NIMH Training Grant No. MH-08258. The authors wish to thank Harold Delaney and Raanan Lipshitz for the computer programs used in the analyses.

<sup>2</sup> Now at Research Triangle Institute. Reprint requests should be addressed to Dr. Anthony J. Conger, Center for Educational Research and Evaluation, Research Triangle Institute, P.O. Box 12194, Research Triangle Park, N. C. 27709.



diagnose specific deficiencies (Kallos, Grabow, and Guarino, 1961) and to assess various psychological traits such as anxiety and neuroticism, (Rashkis and Welsh, 1964). This is partly due, one suspects, to the "clinical tradition" as well as the availability of numerous subscale scores on instruments such as the Wechsler scales. Some of the diagnostic applications are based on criterion group comparisons while others are based on factor analyses (e.g., Lutey, 1966). Based on their own literature survey, as well as relying heavily on the work of Lutey (1966), Robb, Berndardon and Johnson, (1972) present a broad comprehensive approach to the diagnostic use of WISC and WAIS subscale profiles which may become the "standard" for diagnostic applications; however, in order to make valid discriminations among individuals (or groups) one must first confront the problem of profile reliability. Unfortunately, Robb et al. have not considered the overall reliability of the WISC subscales, nor the reliability of the differences of the separate profile dimensions which underlie their approach. In fact, most uses as well as users of profiles have ignored the issue of the reliability of profile differences (perhaps with good reason, in that techniques for the evaluation of profile reliability have not been generally available).

Lutey (1966) and Robb et al., (1972) derived WISC and WAIS profiles based on various factor analyses of the subscales of these instruments. Their method of forming profiles presents problems that are not readily resolvable. First, the scores that are formed are frequently overlapping (but are not weighted so as to be independent). This results not only in highly correlated factor scores, but also results in correlated errors of measurement. For example, for the 7½ year age group, one factor score is formed from the sum of the WISC Information, Arithmetic, and Vocabulary subscales; whereas a second factor score is formed from the sum of Information, Arithmetic, Vocabulary and Comprehension subscales. The only valid difference between these two factor scores would be due to Comprehension! A second problem is that because of the large number of factor scores provided by the Robb et al. approach (9 common factors), strong dependencies exist among the overall set of factor scores. For example, for the 13½ year age group, their "G" factor score can be derived from the sum of their verbal comprehension factor and anxiety (or numerical) factor minus their fluency factor ( $G = VC + A:N - F$ ). Hence, although Robb et al. obtain nine separate factor scores for this age group, there are *at most* eight independent scores. One consequence of these aforementioned problems is that all possible differences among their factor scores might be less reliable than desired. More deleterious however, is that because of linear dependencies and correlated errors of measure-

ment, the actual reliability of their scores can not be unambiguously ascertained.

Based on recent statistical developments (Conger and Lipshitz, 1971, 1973; Conger, 1974) a general measure of profile reliability is available and can be used to establish the reliability of all possible profile (or subscale) comparisons and can also be used to establish a set of independent and maximally reliable profile dimensions. This technique is not unlike the approach discussed by Bock (1966), and its application to instruments such as the WISC is recommended by Cronbach, Gleser, Nanda, and Rajaratnam, (1972). A measure of overall profile reliability is useful for establishing the level of confidence one can have in stating that any observed differences between the profiles of two individuals is a reliable difference. More importantly, however, the most reliable subscale composites would probably provide a better basis than factor analysis for general diagnostic applications because the former approach eliminates unreliable information but retains reliable specific factor information. For the discrimination of clinical subgroups actual validity studies are, of course, requisite; however, to the degree that validity is limited by reliability, the most reliable composites would again provide a better basis for such studies. The purpose of this paper is to determine maximally reliable profile dimensions for the WISC which could be used in various validity studies, and to provide a measure of WISC profile reliability that can serve as a guide for the clinical interpretation of profile differences.

### *Method of Analysis*

Overall profile reliability can be determined for either of two general approaches to profiles. One method for handling profile differences is Cronbach's  $D^2$  and another, that is generally accepted, is Mahalonobis'  $D^2$ . The former profile differences are easier to calculate but are less tractable, in a statistical sense, than the latter. If we let  $X_i$  represent a vector of scores for person  $i$  and  $X_j$  represent a vector of scores for person  $j$  (or a "target" profile) the Cronbach distance measure is

$$D_{ij}^2 = (X_i - X_j)'(X_i - X_j)$$

or in terms of the individual subscales,

$$D_{ij}^2 = \sum_{k=1}^K (X_{ik} - X_{jk})^2$$

where  $X_{ik}$  is the score of person  $i$  on subscale  $k$  (there are  $K$  subscales). The Mahalonobis distance measure for the same scores is found from:

$$D_{ij}^2 = (X_i - X_j)' \Sigma_{xx}^{-1} (X_i - X_j)$$

where  $\Sigma_{xx}^{-1}$  is the inverse of the covariance matrix for the entire set of individuals. In terms of subscale scores,

$$D_{ij}^2 = \sum_{k=1}^K \sum_{k^*=1}^K (X_{i,k} - X_{j,k})(X_{i,k^*} - X_{j,k^*}) \sigma_x^{kk^*}$$

where  $\sigma_x^{kk^*}$  is the  $kk^*$ th element in  $\Sigma_{xx}^{-1}$ . If standardized scores are used, a correlation matrix is employed instead of a covariance matrix, but the same  $D^2$  value will be obtained.

Profile reliability for the Cronbach distances is simply the average of the univariate reliabilities (Conger and Lipshitz, in press):

$$\rho_D^2 = \frac{1}{K} \sum_{k=1}^K \rho_k \quad (1)$$

where  $\rho_k$  is the univariate reliability for subscale  $k$ . Profile reliability for the Mahalonobis distances is found from

$$\rho_D^2 = \frac{1}{K} \text{Trace} (R_{ii}^* R_{xx}^{-1}) \quad (2)$$

where  $R_{xx}^{-1}$  is the inverse of the correlation matrix among the subscales and  $R_{ii}^*$  is the correlation matrix  $R_{xx}$  with univariate reliabilities substituted into the diagonal (Conger and Lipshitz, 1971; 1973).

Although Conger and Lipshitz (1973) have shown that the Cronbach distances are always more reliable than Mahalonobis distances (unless all variables are uncorrelated, in which case both reliabilities are equal), they caution that the choice of these profile distance measures depends on the desired use. The Cronbach distances emphasize common factors of the subscales whereas the Mahalonobis distances allow more weight for independent contribution, but simultaneously allow more weight for unreliable subscales. The real strength of the Mahalonobis approach is that a reduced set of dimensions (or composites) can be found which are uncorrelated with one another and which have maximum reliability.

These maximally reliable and uncorrelated profile dimensions are found by solving the eigenvalue-eigenvector equation

$$R_{ii}^* V_j = \gamma_j^2 R_{xx} V_j \quad (3)$$

where  $V_j$  is termed a canonical vector and provides the weights for the  $j$ th dimension and  $\gamma_j^2$  is a canonical root and equals the reliability of the  $j$ th dimension. The canonical dimension is simply a weighted combination of the initial scores, that is, a new score  $Y_{ij}$  is formed for each individual  $i$  using each canonical vector  $V_j$  as follows:

$$Y_{ij} = \sum_{k=1}^K X_{ik} V_{kj}$$

In fact, and this will be used below, any composite (or dimension) score is simply a weighted combination of the observed scores. Thus, for example, to form the standard verbal-performance contrast, the five verbal subscales are weighted "+1" and the five performance subscales are weighted "-1"; any other scale (e.g., Digit Span) is weighted "0." The verbal-performance contrast is thus:

$$Y_{i(v-p)} = \sum_{k=1}^K X_{ik} W_{vj}$$

where

$W_{kj} = +1$  if subscale  $k$  is a verbal subscale,

$W_{kj} = -1$  if subscale  $k$  is a performance subscale,

and

$W_{kj} = 0$  if subscale  $k$  is neither.

The reliability of any composite scale can be found by:

$$\rho_{ei} = \frac{W_c R_{ii}^* W_c}{W_c' R_{zz} W_c} = \frac{\sum_{k=1}^K \sum_{h=1}^K W_{kc} W_{hc} r_{kh}^*}{\sum_{k=1}^K \sum_{h=1}^K W_{kc} W_{hc} r_{kh}} \quad (4)$$

The canonical profile dimensions have the additional characteristics that if the  $\gamma_j^2$  are ordered according to magnitude (from larger to smaller),  $Y_{i1}$  is the most reliable composite score possible,  $Y_{i2}$  is the most reliable composite score that is uncorrelated with  $Y_{i1}$ ,  $Y_{i3}$  is the most reliable composite score uncorrelated with both  $Y_{i1}$  and  $Y_{i2}$ , etc. Using these facts, a profile user can utilize whatever number of composites deemed by him to have sufficient reliability. Statistical tests for the reliabilities are also available if a test-retest or parallel form approach is adopted (see Bock, 1966 or Conger, 1974). If only the first " $n$ " dimensions are used, then the overall reliability for these dimensions is

$$\rho_{Dn}^2 = \frac{1}{n} \sum_{i=1}^n \gamma_i^2 \quad (5)$$

If all canonical composites are retained, then

$$\rho_{Dn}^2 = \rho_D^2.$$

If the canonical composites are weighted according to their reliability

(more weight for the more reliable dimensions) then an alternative formula, due to Bock (1966), is available:

$$\rho_n = \frac{\sum_{i=1}^n \frac{\gamma_i^2}{1 - \gamma_i^2}}{\sum_{i=1}^n \frac{1}{1 - \gamma_i^2}} \quad (6)$$

Bock's method of weighting yields profile dimensions which have equal standard errors of measurement, thereby allowing easier visual contrasts of profiles and is recommended for this purpose (see Conger, 1974).

### *Interpretation and Rotation of Canonical Dimensions*

While it is very tempting to assign psychological meaning to the canonical dimensions directly from the canonical vectors obtained from equation (2), it should be noted that these vectors contain the weights by which canonical scores are computed from the original variables and do not necessarily give the degree and pattern of association between the canonical dimensions and the original variables. That is, the vectors resemble a factor pattern and not a factor structure (Harman, 1967). The more appropriate matrix for interpretation is the matrix of intercorrelations between the original variables and the canonical variables.

Intercorrelations between the canonical and original variables are found quite simply from the equation

$$R_{xy} = R_{xx}V \quad (7)$$

if the  $y$ 's are standardized canonical dimensions (Conger's weights); however,  $R_{xy}$  is invariant for the various weightings discussed above.

As in factor analysis, the matrix of canonical "loadings" given in equation 7 might be simplified by rotation to a simple or "simpler" structure; however, the dimensions as they are extracted (equation 3) already possess two desirable characteristics: (a) they are sequentially maximally reliable, and (b) they yield uncorrelated total, true and error scores (within each set). Any rotation, beyond a simple change in scale using a diagonal transformation matrix, will destroy both of these properties to some degree. Whether the loss of these properties can be offset by a gain in interpretability can only be decided for each analysis on an individual basis.

Transformations of the maximally reliable dimensions ( $Y$ 's) can be accomplished by an orthonormal rotation matrix  $Q$  ( $Q'Q = I$ ),

$$Z = YQ \quad (8)$$



where the  $Z$ 's are rotated canonical scores. The rotated variables have a structure given by

$$R_{xz} = R_{xy}Q$$

and a pattern

$$W = VQ.$$

In this form the  $Z$ 's are uncorrelated,

$$R_{zz} = Q'V'R_{zz}VQ = I$$

and the rotated true scores have a variance covariance matrix given by:

$$R_{T_zT_z} = Q'V'R_{T_T}VQ = Q'\Lambda Q$$

where  $\Lambda$  is a diagonal matrix of unrotated canonical reliabilities. It should be noted that the trace of  $R_{T_zT_z}$  is merely the sum of the unrotated canonical reliabilities; however,  $R_{T_zT_z}$  is not diagonal (unless  $Q$  is). The multivariate reliability in the rotated space does, however, remain invariant as can be shown by substituting  $R_{T_zT_z}$  and  $R_{zz}$  into equation (2).

### *Data*

Data have been taken from the WISC manual (Wechsler, 1949). Wechsler provides correlation matrices for 200 subjects at each of three age levels: 7½, 10½, and 13½ year olds. Wechsler also provides split-half reliability estimates for eleven of the twelve subscales for each age group separately (Coding, since it is a speeded test, can not be used). Only the eleven subscales for which reliability estimates are available are used in the following analyses.

### *Analyses*

Univariate reliabilities and intercorrelation matrices were substituted into equations (1), (2) and (3) to obtain overall estimates of profile reliability. The vectors obtained from equation (3) were then substituted into (7) to obtain the unrotated canonical structure. This structure, for each age group separately, was then subjected to a row normalized varimax rotation on the retained dimensions.

### *Results: Canonical Reliability*

If profile differences are calculated by the Cronbach method, the overall reliabilities are .68, .76, and .75 for the 7½, 10½ and 13½ age

groups respectively (equation 1). These overall reliabilities do not compare well with the total test reliabilities of .92, .95, and .94 (Wechsler, 1949) but are quite reasonable in magnitude compared to the reported verbal-performance difference reliabilities of .68, .76, and .84 as found from equation 5 for total Verbal versus total Performance.

The Mahalanobis difference reliabilities (equation 2) are .49, .51 and .53 for the respective age groups. These values are substantially lower than the Cronbach difference reliabilities and indicate that all possible comparisons should not be made from the WISC subscales, but a dimensional analysis might provide useful information as to which kinds of profile comparisons might be made.

### *Results: Canonical Dimensions*

The roots obtained from equation 3 indicate that there are, at most, five reasonably reliable dimensions, and more definitely, two very reliable dimensions within each age group. In all three groups there was a break in the eigenvalues after the first two roots and a somewhat smaller break after the first five (Table 1). While the reliabilities of (unrotated) dimensions 3, 4, and 5 might fall below a satisfactory magnitude for some test users (e.g., the reliability of dimensions 5 for 7½ year olds is only .55) all five dimensions have been retained for interpretation and rotation. The overall reliabilities of the first five dimensions are .71, .75, and .77 if equated on total variance or .83, .90 and .88 if equated on error variance.

### *Interpretations: Unrotated Dimensions*

The first unrotated dimension (Table 2) in all three age groups has high positive correlations with all WISC subscales. This pattern of

TABLE 1.  
*Canonical Vector Reliabilities for WISC*

Root	7½ Year Olds	10½ Year Olds	13½ Year Olds
1	.94	.97	.96
2	.82	.84	.86
3	.65	.71	.73
4	.60	.66	.69
5	.55	.59	.60
6	.44	.54	.51
7	.43	.46	.49
8	.37	.41	.36
9	.32	.27	.33
10	.19	.13	.29
11	.09	.08	.06

TABLE 2  
Correlations\* of Unrotated Dimensions with WISC Subscales

Scale	7½ Year Olds					10½ Year Olds					13½ Year Olds				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Information	65	35	-08	-36	-24	84	25	05	09	16	85	24	07	-13	-09
Comprehension	59	25	-06	08	42	76	18	-31	-09	-10	68	32	-08	-19	28
Arithmetic	59	35	04	-08	-31	77	20	49	11	-05	64	18	48	-37	09
Similarities	61	29	07	-24	-23	74	38	21	01	20	80	18	01	00	31
Vocabulary	71	46	-15	-17	28	91	18	-30	-04	-08	86	33	-26	16	-20
Digit Span	54	33	11	32	-10	54	03	07	20	-58	46	09	19	-16	-11
Picture Completion	50	16	00	35	44	55	-23	-18	-10	59	53	-43	03	08	09
Picture Arrangement	69	07	38	43	-19	66	-06	02	-18	-08	56	-21	-07	42	54
Block Design	74	-48	-44	08	-08	73	-55	16	-29	-02	72	-60	-10	-24	-08
Object Assembly	65	-31	13	17	-11	53	-46	01	-11	-17	54	-61	-04	25	-11
Mazes	61	-53	43	-31	18	59	-51	-10	60	07	49	-07	67	43	-20
Reliability	94	82	65	60	55	97	84	71	66	59	96	86	73	69	60

\* Decimal points not shown

loadings indicates that the most reliable comparison among individuals is a simple comparison of profile levels (i.e., total test score). Although the weights for the various subscales (Table 3) are not equal, the reliability of the first canonical composite is not sufficiently different from total test reliability to warrant differential weighting of the subscales.

The second most reliable dimension (Table 2) closely resembles a verbal-performance contrast with some differences in pattern (Table 3) and structure (Table 2) across age groups. On first appearances, the signs for the 7½ year age group do not seem to conform to a simple dichotomy of verbal and performance scales; however, they do match the results of previous factor analyses for this age group (cf. Cohen, 1959). In all three groups, the reliability of this dimension exceeds that of the standard verbal-performance difference and should probably be used in its place.

The remaining three dimensions have different patterns and structures (Table 3) across age groups and at best, a tenuous interpretation could be offered, but in the interest of brevity the reader will be spared. A more detailed interpretation is given by Conger and Conger (1972).

### *Interpretations: Rotated Dimensions*

The results of the row normalized varimax are given in Tables 4 and 5. The resulting structures of the various age groups were compared for factor similarity by using Tucker's  $\phi$  coefficient of factor congruence (Harman, 1967, p. 270). A good across age group fit was obtained for four of the five rotated dimensions; furthermore, these four dimensions (1-4, Table 4) showed similar patterns of interrelationships both across and within age groups when cast into a

TABLE 3  
*Canonical Weights for First Two Maximally Reliable WISC Dimensions*

Scale	First Dimension			Second Dimension		
	7½	10½	13½	7½	10½	13½
Information	.12	.13	.13	.19	.20	.19
Comprehension	.09	.09	.09	.11	.11	.16
Arithmetic	.10	.15	.11	.17	.20	.11
Similarities	.11	.12	.14	.16	.32	.12
Vocabulary	.20	.32	.32	.37	.32	.46
Digit Span	.09	.04	.04	.15	.01	.03
Picture Completion	.08	.05	.06	.07	-.11	-.19
Picture Arrangement	.16	.07	.08	.05	-.04	-.11
Block Design	.30	.18	.23	-.55	-.68	-.70
Object Assembly	.11	.05	.07	-.16	-.20	-.29
Mazes	.19	.10	.07	-.47	-.43	-.04

TABLE 4  
*Correlations\* of Rotated Dimensions with WISC Subscales*

Scale	1st Dimension			2nd Dimension			3rd Dimension			4th Dimension			5th Dimension		
	7½	10½	13½	7½	10½	13½	7½	10½	13½	7½	10½	13½	7½	10½	13½
Information	82	67	65	16	51	55	14	17	23	09	21	14	07	11	12
Comprehension	24	29	60	71	76	41	11	23	04	11	10	-14	13	00	37
Arithmetic	67	86	87	11	18	07	09	28	10	01	14	21	33	-17	07
Similarities	69	77	56	13	39	40	07	09	20	15	03	05	21	12	50
Vocabulary	61	40	33	68	83	88	08	29	20	05	18	11	03	01	19
Digit Span	33	28	45	32	35	21	05	27	15	-07	24	15	55	-60	-04
Picture Completion	00	22	16	69	29	07	12	32	60	06	25	14	30	67	27
Picture Arrangement	24	33	06	21	40	14	14	46	35	22	05	14	82	00	81
Block Design	19	25	33	19	18	14	92	90	90	22	18	-06	12	12	05
Object Assembly	15	06	-04	11	20	15	44	66	78	39	24	25	44	-04	19
Mazes	13	14	25	11	16	10	23	32	20	94	91	89	11	04	12
Reliability	75	82	79	66	81	81	76	82	87	72	72	71	66	60	66

\* Decimal points not shown.



TABLE 5  
*Canonical Weights for Rotated Dimensions*

Scale	1st Dimension			2nd Dimension			3rd Dimension			4th Dimension			5th Dimension		
	7½	10½	13½	7½	10½	13½	7½	10½	13½	7½	10½	13½	7½	10½	13½
Information	.52	.30	.31	-.20	.03	.16	.00	-.19	-.04	-.01	.09	.03	-.17	.23	-.22
Comprehension	-.12	-.16	.25	.47	.36	-.04	-.06	-.02	-.12	.02	-.08	-.27	-.09	-.06	.26
Arithmetic	.34	.75	.67	-.23	-.47	-.36	-.02	.02	-.12	-.09	.00	.09	.14	.37	-.10
Similarities	.41	.51	.22	-.21	-.11	-.10	-.07	-.19	-.12	.04	-.11	-.14	.00	.30	.52
Vocabulary	.24	-.37	-.44	.62	1.04	1.19	-.12	-.11	-.04	-.01	-.06	.08	-.39	-.13	-.27
Digit Span	.00	-.04	.12	.05	.07	-.01	-.04	.06	.00	-.15	.08	.05	.36	-.60	-.11
Picture Completion	-.29	.03	-.04	.51	.00	-.09	-.04	-.03	.16	-.04	.07	.04	.11	.75	.14
Picture Arrangement	-.12	-.01	-.18	-.10	.07	-.15	-.10	.18	-.01	-.01	-.15	.01	.83	-.05	.87
Block Design	-.06	-.07	.18	-.05	-.27	.14	1.14	.99	.82	-.29	-.34	-.41	-.18	.14	-.42
Object Assembly	-.07	-.11	-.24	-.09	.00	.05	.11	.24	.29	.09	-.01	.10	.26	-.12	.02
Mazes	-.03	-.06	-.03	.01	-.14	-.08	-.27	-.20	-.02	1.08	1.14	.96	-.19	-.02	-.07

multitrait-multimethod (age groups) matrix (not shown). The intercorrelations of the rotated true scores are very near zero with average absolute values of .06, .04 and .04 and absolute maximum intercorrelations of .14, .11 and .10 for the respective 7½, 10½ and 13½ year age groups.

The first rotated dimension (Table 4) corresponds closely to a factor of Verbal Comprehension (Robb, Berndardoni, and Johnson, 1972) and has across age groups  $\phi$ 's of .96, .91, and .94. The differences across age groups primarily involve differences in the Comprehension subscale (agreeing with factor analytic results) and Vocabulary (not agreeing with factor analytic results). The increasing loss of importance of Vocabulary for increases in age is, however, offset by the second rotated dimension.

The second dimension has loadings somewhat similar to the first, especially for the younger age groups. The  $\phi$  values of factor congruence are .90, .76 and .94. The loadings correspond somewhat to the "R" factor of Relevance in the 7½ and 10½ year olds, (Robb, et al.) which has high loadings for Comprehension and Vocabulary. Vocabulary tends to dominate the factor, but the other verbal scales also have high loadings. The R and Verbal Comprehension factors seem to be somewhat confounded in the 13½ year age group.

The third dimension has good agreements across age groups ( $\phi$ 's of .92, .90, and .94) and, especially for 13½ year olds, resembles a Perceptual Organization factor (Robb, et al.). In all three groups, this factor is dominated by Block Design (loadings of .92, .90 and .90). Deviations from the Perceptual Organization factor are primarily due to the absence of Mazes (see dimension 4) in all groups and low loadings for Picture Arrangement in 7½ year olds and Picture Completion in 10½ year olds (see dimension 5 for both groups).

The fourth rotated dimension is a very clear specific factor, having a high loading for Mazes (.94, .91, and .89) with good agreement across age groups ( $\phi$ 's of .89, .87 and .93). Mazes contributes very independent information to the rotated profile dimensions and thus could be added or subtracted to other dimensions with no loss in information. For example, a better Perceptual Organization factor score could be obtained by adding scores on dimension 4 to scores on dimension 3.

The fifth rotated dimension shows some correspondence between 7½ and 13½ year olds ( $\phi = .76$ ) but little agreement for other pairings of the age groups (-.13 for 7½ versus 10½ and .26 for 10½ versus 13½). In the 7½ year olds, this dimension resembles the Freedom from Distractibility factor but has little resemblance to that factor or any other major factor of 10½ and 13½ year olds. This 'residual' dimension has the lowest reliability of the rotated dimensions for each age

group, but its deletion would result in only moderate gains in multivariate reliability (.01, .04, and .03 for the respective age group).

### *Comparisons of Rotated and Unrotated Dimensions*

The rotated dimensions (Table 4) do possess a reasonable simple structure and across age groups match with little loss in the independence of the underlying true scores. There is, however, a leveling out of the reliabilities of the rotated dimensions with respective maximum values of .76, .82 and .87.

The unrotated solutions have only two dimensions and these are general factors which correspond closely to the major uses of the WISC scores, i.e., overall IQ and Verbal-Performance differences. The reliabilities of each of these two dimensions exceed the maxima of the rotated dimensions (with the exception of the Verbal Performance contrast versus the Perceptual Organization factor in 13½ year olds, i.e., .86 versus .87).

There seem to be advantages to both types of solutions and perhaps a choice between them should be predicated on the uses to which the WISC profile will be subjected. Clinical applications would seem to favor retaining only the two major unrotated dimensions while research purposes might be better served by the rotated dimensions. If one further considers that the Mazes subscale is frequently omitted during WISC testings, the advantages of the unrotated solution seem enhanced. It is possible, of course, to keep the rotated dimensions and to combine them as needed to form the level and Verbal-Performance contrasts.

### *A Note on Unreliable Dimensions*

In the same manner that the most reliable dimensions indicate what reliable comparisons can be made among individuals, the least reliable dimensions indicate the comparisons that should not be made. In particular, the tenth dimension for 7½ and 13½ year olds and the eleventh dimension for 10½ year olds (Table 6) indicates that a composite contrasting Information with Comprehension plus Arithmetic is very unreliable ( $\rho = .19, .08$  and  $.29$  respectively). The eleventh composite for 7½ and 13½ year olds is also similar, with the pattern of weights indicating a possible simultaneous contrast among verbal scales (especially Information versus Vocabulary) and among performance scales. To the extent that the verbal scales have high common variance, differences among them would be very unreliable (the same would be true for performance scales). The tenth dimension for 10½ year olds involves contrasts among the three least reliable subscales:

TABLE 6  
*Canonical Weights for Least Reliable WISC Dimensions*

Scale	7½ Year Olds		10½ Year Olds		13½ Year Olds	
	$V_{10}$	$V_{11}$	$V_{10}$	$V_{11}$	$V_{10}$	$V_{11}$
Information	1.05	.40	.28	1.40	1.42	.63
Comprehension	-.45	.68	.07	-.97	-.89	.46
Arithmetic	-.76	-.25	-.20	-.73	-.31	-.28
Similarities	-.30	-.13	-.29	-.13	.20	-.07
Vocabulary	-.11	-.54	-.45	.04	-.51	-.70
Digit Span	-.01	.63	.96	.27	-.02	.56
Picture Completion	.46	-.30	.70	-.07	.11	-.62
Picture Arrangement	.23	-.67	-.24	.16	.27	-.04
Block Design	-.06	-.31	.40	-.25	-.11	-.61
Object Assembly	-.05	.81	-.77	.58	-.26	1.18
Mazes	-.03	-.11	-.25	-.27	-.23	-.24
Reliability	.19	.09	.13	.08	.29	.06

Object Assembly versus Digit Span and Picture Completion. Of particular interest is that because these three subscales have very low correlations with the remaining scales, one would expect that large differences in profile patterns are *a priori* more probable, but simultaneously, less reliable.

### Summary

The preceding results lead to the following recommendations concerning the use of the WISC subscales for profile comparisons:

1. As far as "clinical" use of WISC profile comparisons is concerned, our advice is "caveat emptor." Pure intuition as to whether any two profiles differ or not will quite frequently capitalize on an unreliable difference. If profiles are compared by the Cronbach  $D^2$  method, the reliability of the differences will be good (around .73) but not outstanding. If all possible differences are allowed, the reliability of an "average" difference is quite unacceptable (around .51).
2. While five uncorrelated dimensions are probably sufficient to establish a set of reasonably reliable WISC profile dimensions as predicted by Cronbach (Cronbach et al., 1972); only two dimensions have an unambiguous and readily interpretable pattern. The clinician should probably restrict himself to diagnoses using these two largest dimensions (total IQ and Verbal Performance differences) until satisfactory reliability and validity are established for the other uncorrelated dimensions.

It is also possible to establish four dimensions which are similar across age groups and which correspond to factor analytic dimensions. Three of these dimensions correspond to group factors of Verbal Comprehension, Relevance and Perceptual Organization while the

fourth is specific to Mazes. Thus, reliable individual differences can be found for these three group factors and the Mazes subscale can be considered as providing reliable individual differences beyond that. These four dimensions might be quite useful for research purposes involving the WISC subscales when it is not considered feasible to utilize all 11 subscales (e.g., for Multiple Regression or Analysis of Variance analyses). The advantage of these new dimensions over the subscales themselves and over factor scores is that these dimensions are all reasonably reliable, reasonably interpretable and are virtually independent. Such a claim could not be made for the numerous factor scores discussed by Robb et al. (1972) or for the separate subscales themselves.

3. The most reliable dimension for all three age groups involves approximately a simple sum of the subscale scores. While Vocabulary and Block Design do have relatively high weights, the gain in reliability of the more complex sum is insignificant by comparison. The second most reliable dimension is a Verbal-Performance contrast and in this case, two features should be noted. The weighted Verbal-Performance contrast does provide a gain in reliability over a simpler sum, and the weights differ for the different age groups.

4. The least reliable dimensions indicate that comparisons should not be made among subscales that are highly correlated relative to their level of reliability. In particular, Information versus Comprehension plus Arithmetic is a source of very unreliable individual differences.

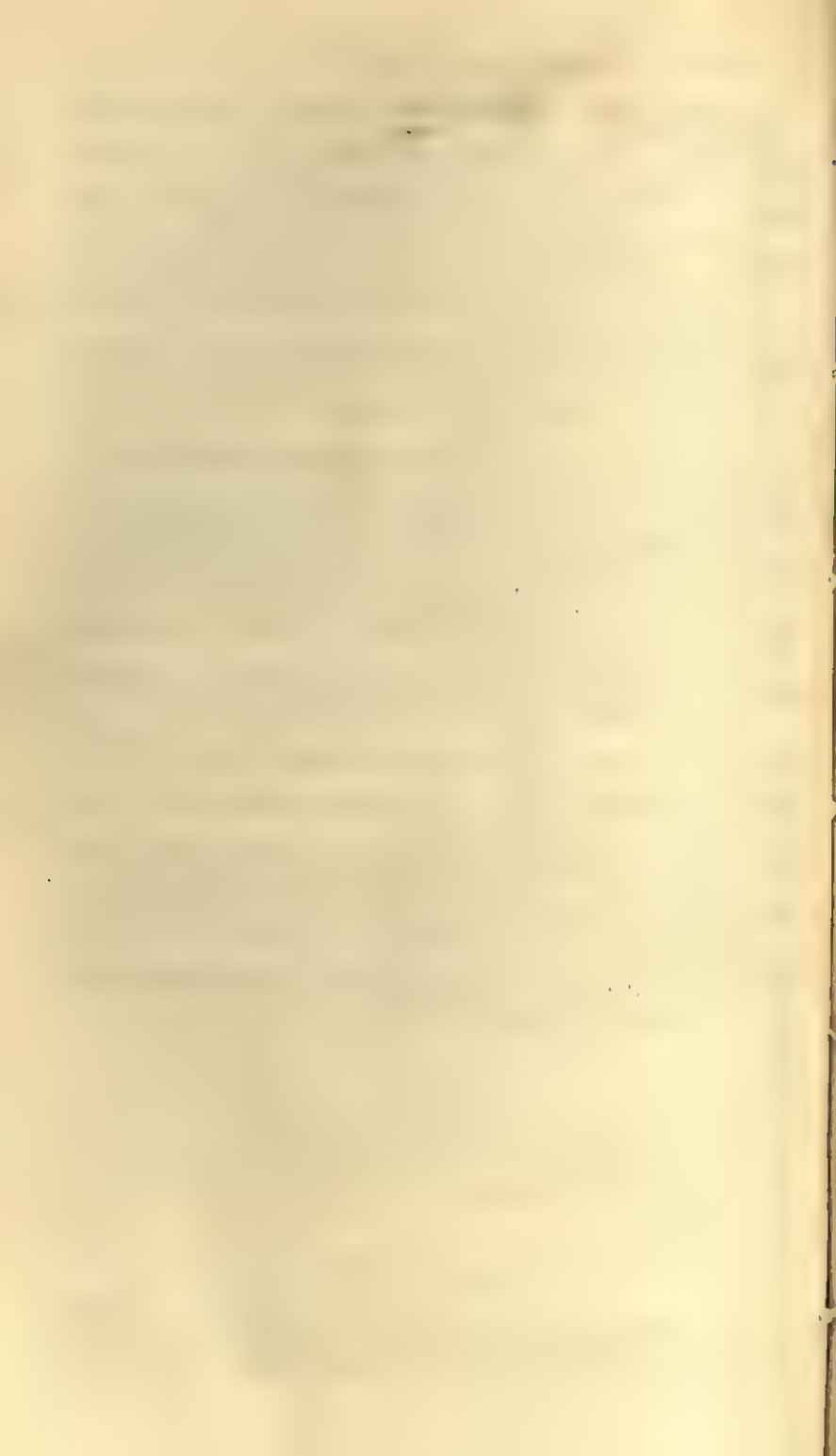
Based on the above results, a general statement seems warranted. It is no great surprise that the WISC functions best for the purpose for which it was designed, i.e., large amounts of time and money were expended to develop an instrument that would "measure" general ability and verbal and performance abilities. The WISC serves these purposes rather well. What is surprising is that psychologists take such an instrument and combine and permute the subscales in  $n$ -different ways with little concern for such test fundamentals as reliability and validity. There seems to be a type of logic involved that asserts: "if the instrument is good in serving one purpose, it, therefore, must be good in serving any purpose." As Peterson (1968) asserts, albeit in a somewhat different context, "there is no cheap way to study human behavior."

## REFERENCES

- Anastasi, A. *Psychological testing*. New York: Macmillan, 1968.  
Bock, R. D. Contributions of multivariate experimental designs to educational research. In R. B. Cattell (Ed.) *Handbook of Mul-*



- tivariate Experimental Psychology*. Chicago: Rand McNally, 1966, 820-840.
- Cohen, J. The factorial structure of the WISC at ages 7-6, 10-6, and 13-6. *Journal of Psychology*, 1959, 23, 285-290.
- Conger, A. J. Estimating profile reliability and maximally reliable composites. *Multivariate Behavioral Research*, 1974, 9, 85-104.
- Conger, A. J. and Lipshitz, R. Canonical reliability for profiles and test batteries. Chapel Hill, N. C. The L. L. Thurstone Psychometric Laboratory, University of North Carolina, Research Bulletin 97, 1971.
- Conger, A. J. and Conger, J. C. Maximally reliable profile dimensions for the WISC. Chapel Hill, N. C.: The L. L. Thurstone Psychometric Laboratory, University of North Carolina, Research Bulletin 109, 1972.
- Conger, A. J. and Lipshitz, R. Measures of reliability for profiles and test batteries. *Psychometrika*, 1973, 38, 411-427.
- Cronbach, L. J., Gleser, G. C., Nanda, H. and Rajaratnam, N. *The dependability of behavioral measurement*. New York: Wiley, 1972.
- Gainer, W. L. The ability of the WISC subtests to discriminate between boys and girls classified as educable mentally retarded. *California Journal of Educational Research*, 1965, 16, 85-92.
- Harman, H. H. *Modern factor analysis*. Chicago: University of Chicago Press, 1967.
- Kallos, G. L., Grabow, J. H. and Guarino, E. A. The WISC profile of disabled readers. *Personnel and Guidance Journal*, 1961, 39, 476-478.
- Lutey, C. *Individual intelligence testing: A manual*. Greeley, Colorado: Executary, Inc., 1966.
- Peterson, D. *The clinical study of social behavior*. New York: Appleton-Century-Crofts, 1968.
- Rashkis, H. A. and Welsh, G. S. Detection of anxiety by the use of the Wechsler scale. *Journal of Clinical Psychology*, 1964, 20, 354-357.
- Robb, G. P., Berndardoni, L. C. and Johnson, R. W. *Assessment of individual mental ability*. Scranton, Pa.: International Text Book Co., 1972.
- Wechsler, D. *Wechsler intelligence scale for children: Manual*. New York: The Psychological Corporation, 1949.



## PAIRED COMPARISONS INTRANSITIVITY: TRENDS ACROSS DOMAINS OF CONTENT AND ACROSS GROUPS OF SUBJECTS

DARWIN D. HENDEL  
Measurement Services Center  
University of Minnesota

The present study investigated the extent to which intransitivity, as measured by the total circular triad (TCT) score in the method of paired comparisons, varies across different domains of content and across different groups of subjects. Three paired comparisons instruments were administered to 276 high school students and to 358 college students. Results indicated statistically significant ( $p \leq .001$ ) correlations among the three TCT indices for both groups of subjects.

THE method of paired comparisons, one application of Thurstone's (1927) "law of comparative judgment," can be used to obtain preferences for a set of stimulus objects. In addition, the method yields an index of response intransitivity which has been termed the "total circular triad" score.

One of the aspects of the total circular triad score which has been questioned concerns the generality of response intransitivity as measured by the total circular triad score. Does intransitivity occur across instruments or is intransitivity instrument specific? Do relationships among intransitivity indices for one group of subjects replicate for other groups of subjects?

Although previous investigations (Gulliksen, 1964; Ace and Dawis, 1972) have obtained intransitivity indices for more than one instrument, the generality of intransitivity across subject groups has not been established. The purpose of the present study was to investigate the generality of intransitivity across widely different domains of content and across different groups of subjects.

*Method*

Two groups of subjects were used in the present investigation. The first group consisted of 276 students from Cooper Senior High School in the Robbinsdale (Minnesota) school district; the students participated in the questionnaire session as part of the class work in vocational education. The mean age was 17.1 years ( $SD = .67$ ); 96 (35%) of the students were males and 180 (65%), females. The second group consisted of 358 students enrolled in Introductory Psychology at the University of Minnesota; students received experimental points for participating in the questionnaire session. The mean age was 19.9 years ( $SD = 2.48$ ); 168 (47%) of the students were males and 190 (53%), females.

Three paired comparisons instruments were used to obtain the total circular triad indices of response intransitivity. Each instrument contained 20 statements in a complete paired comparisons format which resulted in a total of 190 paired comparisons items [ $n(n - 1)$  items where  $n$  equals the number of statements]. The first instrument used was the Minnesota Importance Questionnaire (MIQ; Gay, Weiss, Hendel, Dawis and Lofquist, 1971) which contains statements of vocational needs (e.g., "I could do something that makes use of my abilities"). The second instrument, designed by the author, was the Mate Selection Questionnaire (MSQ), which contains qualities presumed to be used frequently in choosing a mate (e.g., "physical attractiveness"). The third instrument, designed by the author, was the Food Preference Questionnaire (FPQ) which obtained preferences for main course meals (e.g., "hot beef sandwiches"). For each instrument, the subject indicated his/her preference between pairs of responses.

Total circular triad scores were calculated according to Kendall's (1955, p. 125) formula:

$$TCT = 1/6n(n - 1)(2n - 1) - \sum_{i=1}^n X_i/2$$

where the  $X_i$  represent an individual's scale scores on the  $n$  stimulus objects. The 20 scale scores for each of the three inventories reflected the number of times each of the stimulus objects was preferred in the total set of 190 items. Scale scores could range from 0 (statement was never chosen in any pair) to 19 (statement was chosen every time it appeared in a pair).

Pearson product-moment correlations (Guilford, 1956, p. 95) were calculated among the TCT scores on the MIQ, MSQ and FPQ for the high school and college student groups separately.

TABLE 1

*Intercorrelations of Total Circular Triad Scores on the MIQ, MSQ, and FPQ for High School and College Student Groups*

Group	Intercorrelations <sup>a</sup>			<i>X</i>	<i>S.D.</i>
High School Group ( <i>N</i> = 234)	TCT on MIQ	TCT on MSQ	TCT on FPQ	91.27	71.20
TCT on MIQ				61.16	49.26
TCT on MSQ	.58***			34.18	35.97
TCT on FPQ	.34***	.39***			
College Student Group ( <i>N</i> = 348)	TCT on MIQ	TCT on MSQ	TCT on FPQ	52.39	40.42
TCT on MIQ				35.34	27.65
TCT on MSQ	.39***			24.06	23.42
TCT on FPQ	.25***	.33***			

<sup>a</sup> Minimum Pearson product-moment correlation coefficients necessary for significance at the .05, .01, and .001 levels of significance are .127, .166, and .210 respectively for the high school group (232 *df*) and .105, .138, and .174 respectively for the college student group (346 *df*).

\*  $p \leq .05$ .

\*\*  $p \leq .01$ .

\*\*\*  $p \leq .001$ .

### *Results and Discussion*

Table 1 contains the Pearson product-moment correlations among the three TCT scores for both subject groups. The correlations were significant at  $p \leq .001$  for both the high school and college student groups. Although the level of the correlations was consistently higher for the high school group, the pattern in the correlation matrix was similar for both groups.

Considering the TCT scores as indices of response intransitivity, the obtained correlations suggest that the tendency to respond intransitively does generalize across different domains of instrument content and across different groups of subjects.

The results suggest that intransitivity is not a random error variable. Intransitivity can be considered as an indicator of true differences among individuals rather than as an instrument specific index of error in responding to paired comparisons instruments. However, intransitivity is not totally generalizable from one content domain to another. Although an individual who is intransitive on one instrument is more likely to be intransitive on other instruments, specific sets of stimuli probably interact with the individual's basic tendency to respond intransitively.

### REFERENCES

- Ace, M. E. and Dawis, R. V. The contributions of questionnaire length, format and type of score to response inconsistency.



EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1972, 32, 1003-1011.

Gay, E. G., Weiss, D. J., Hendel, D. D., Dawis, R. V. and Lofquist, L. H. Manual for the Minnesota Importance Questionnaire. *Minnesota Studies in Vocational Rehabilitation*, 1971, 28.

Guilford, J. P. *Fundamental statistics in psychology and education*. New York: McGraw-Hill, 1956.

Gulliksen, H. Intercultural studies of attitudes. In Frederiksen, N. and Gulliksen, H. (Eds.) *Contributions to mathematical psychology*. New York: Holt, Rinehart and Winston, 1964.

Kendall, M. G. *Rank correlation methods*. New York: Hafner, 1955.

Thurstone, L. L. A law of comparative judgment. *Psychological Review*, 1927, 34, 273-286.

## AN ANALYSIS OF THE MEANING OF THE QUESTION MARK RESPONSE CATEGORY IN ATTITUDE SCALES<sup>1</sup>

BERNARD DuBOIS<sup>2</sup> AND JOHN A. BURNS

Northwestern University

Although most scaling formats include an intermediate or neutral response category, little research has been devoted to the analysis of the meaning *respondents* attach to this category. Results obtained from ten different scales, across two types of item formats (Likert and Polar Choice) support the traditional method of scoring the "?" answer. Although the meaning *respondents* imply when selecting the "?" is not more ambiguous than the meaning implied in the selection of the other response categories, there does exist evidence for the presence of a variety of uses of the "?" including response styles, ambivalence and indifference. Various suggestions are made for further research and alternate methods of approach to the meaning of the question mark response category.

ALTHOUGH a multiple-indicator approach to the assessment of social attitudes has been advocated by various authors (Cook and Selitz, 1970; Webb, Campbell, Schwartz, and Sechrest, 1966; Summers, 1970) most current research still relies on the time-honored self-report technique of attitude scales.

An attitude scale represents an attempt to measure an individual's dispositions toward a given issue by asking him to express his degree

---

<sup>1</sup> The preparation of this paper has been supported by NSF Grants GS1309X and GS30273X.

<sup>2</sup> The authors are deeply grateful to Donald T. Campbell for his numerous and helpful comments and suggestions made at various stages of this research. Also, the authors would like to thank Kelman Kaplan for commenting on an earlier draft of this paper. Bernard Dubois is now at C.E.S.A (Centre d'Enseignement Supérieur des Affaires), 78350 Jouy-en-Josas, France and John A. Burns is at Social Science Unit, Department of the Environment, 1090 W. Pender, Vancouver, Canada.

of agreement or disagreement to a group of statements. Although a variety of scales are found in the literature (e.g., Thurstone Equal Appearing Interval Scale, Guttman Scalogram), the type most frequently employed is the summated rating scale (Likert, 1932), more commonly known as the Likert scale. In a Likert scale, the format of responses to an item traditionally consists of one or more levels of agreement (e.g., "Strongly Agree," "Agree"), one or more levels of disagreement (e.g., "Strongly Disagree," "Disagree") and one level of indecision or indetermination (the question mark ("?"), "undecided" or "neutral" response category).

In the psychometric literature a substantial amount of research has focused on the properties of the entire scale (validity, reliability, etc.), or on the properties of individual items (representativeness, ambiguity, etc.). However, investigators generally do not examine the meaning of each item's response category, regarding these meanings as relatively unambiguous: "Strongly Agree" means that the respondent agrees more strongly than just agreeing; "Strongly Disagree" means disagreeing more strongly than just disagreeing; the neutral point means just that—in the middle. Subsequently, the researcher typically assigns scores implying that the response categories are on an interval scale. For example, he would use the following scoring system: "Strongly Agree" = 5; "Agree" = 4; "?" = 3; "Disagree" = 2; "Strongly Disagree" = 1. It is always assumed that the respondent uses the same set of meanings as the researcher.

Although Likert (1932) had originally presented data which supported this scoring system, more recent research (cf. Stanley and Wang, 1968; Wang and Stanley, 1970) indicates that these assumptions may not be valid. For example, due to the item's wording, respondents may actually be interpreting "Strongly Agree" as being in the region of a score of six or seven rather than five; they see it as being "very Strongly Agree," i.e., more than just one equal interval from "Agree." Various post hoc statistical methods have been developed which generate item response scores which are more sensitive to the meanings respondents attach to their responses (e.g., MacDonald, 1968).

This increased awareness on the part of researchers to the meanings of individual responses is a welcome initiative. Researchers need to consider more than just the score obtained on each item—they need to understand how the subject *interprets* each item and its response alternatives before the difficulties involved in measuring attitudes with pre-coded instruments can be overcome. In general, too many aspects of the individual items of attitude scales are left untested.

The problem of the meaning and appropriate scoring of the "?"

response category is certainly such an aspect. According to the small amount of research conducted on this topic (Cronbach, 1946; Edwards, 1946; Worthy, 1969; Golberg, 1971; and Kaplan, 1972), the essential meaning of the neutral category or question mark is either one of ambivalence or indifference.

The *ambivalent* respondent has "mixed feelings," i.e., positive and negative sentiments concentrated on the same object (Brown, 1965), and cannot make up his mind as to whether he agrees or disagrees with the proposed statement. To express his internal state of indetermination, he marks the "?" answer. Ambivalence is often the result of over-involvement with the issue under analysis. Well familiarized with the pro and con arguments, the ambivalent respondent finds it difficult to make a final choice.

The *indifferent* respondent checks the "?" response because he has minimal concern for the topic involved in the statement. While ambivalence often results from overinvolvement, indifference generally indicates underinvolvement. Operationally, it therefore may be possible to separate ambivalent from indifferent respondents if one has data available concerning the respondents' level of involvement. It is hypothesized that the ambivalent respondent should exhibit a high level of attitude intensity while a low level of intensity should characterize the indifferent respondent (cf. Diab, 1965).

Ambivalence and indifference, however, are by no means the only two factors which account for the use of the "?" response category. Some respondents, for example, may check the "?" mainly because, although not indifferent, they do not feel competent enough or sufficiently informed to take a position. Others might use it as an indirect way of expressing their refusal to reveal their personal feelings. Still others might use it because they do not understand the attitudinal statement.

Various positions have been taken on the scoring of the question mark. At one extreme, some researchers, for example Cronbach (1946), have suggested that given the ambiguous meaning of this category, attitude researchers should probably discontinue its use. Other researchers have adopted the reverse approach by making the meaning of the "?" more *specific*. Goldberg (1971) for example suggests the following format to supplement the "?" answer:

Note: Where you have indicated your feelings about one of the above statements by circling a question mark, please go back and write in one of the following next to the question mark:

- I Indifferent
- M Mixed feelings

DK Don't know

O Other; please specify your feelings

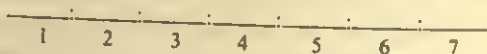
Most attitude researchers, however, seem to have adopted an intermediate and rather intriguing position; although recognizing that respondents may use the question mark response category for a variety of reasons, they code it as if it were an indicator of a middle position within a favorableness-unfavorableness continuum, i.e., as if it represented the arithmetic average of the two positions on either side of it. This position is exemplified by Osgood, Suci, and Tannenbaum (1957), in their instructions for the semantic differential scale:

The direction toward which you check, of course, depends upon which of the two ends of the scale seems most characteristic of the thing you're judging.

If you consider the concept to be *neutral* on the scale, both sides of the scale *equally associated* with the concept, or if the scale is *completely irrelevant*, unrelated to the concept, then you should place your checkmark in the middle space:

safe: \_\_\_\_\_ x \_\_\_\_\_ dangerous

Osgood, Suci and Tannenbaum then suggest the following scoring system:



Clearly, this is equivalent to considering the "neutral" point as an indicator of a middle position along a continuum. Most elementary textbooks on attitude measurement (e.g., Oppenheim, 1966, p. 133) also suggest this approach.

Therefore, it appears that the most popular current practice followed by attitude researchers with respect to the interpretation of the question mark response category rests on the implicit assumption that: either (1) respondents use the question mark response category as an indicator of an intermediate position to a far greater extent than they use it for other reasons so that this particular use is the only one which really deserves consideration; or (2) respondents use the question mark response category for many various reasons but these reasons tend to counterbalance each other so that the final effect is similar to the one obtained had respondents used the question mark response category for the sole purpose of expressing a neutral position.

This paper addresses itself to the problem of testing this double assumption.

If the question mark response category is used by respondents



almost uniquely as an expression of an average position on the agree-disagree continuum, two specific results should be expected.

First, when we plot the respondents' total (all-item) scores against the scores obtained on any item of a scale, the dispersion of the total scores obtained for those respondents who checked the "?" answer on the item under consideration should be similar to the dispersion of the total scores obtained for the respondents checking any other response category. This implies that the respondent's use of the "?" answer is *not more equivocal* than the use of any other response category and therefore the "?" answer represents a given position on the construct being measured in the same manner as do the other response categories (given the assumption that the scale under consideration is internally consistent).

Second, the mean of the distribution of the total scores obtained for those respondents who used the "?" answer should fit the pattern of means suggested by the distributions of total scores corresponding to the other response categories. The reason of course is that the "?" is assumed to indicate an average position on the agree-disagree continuum.

Figure 1 illustrates the above situation: the mean of the distribution of total scores for the question mark category ( $\bar{X}_?$ ) fits the pattern of means (here a straight line) and its dispersion of total scores is similar to the other categories' dispersions of total scores.

When the question mark category is used for other purposes than to indicate an intermediate position, this particular pattern should not be expected. In fact, if we assume that the motivations underlying the use of the question mark category vary among respondents, the observed standard deviation for the question mark category distribution should be significantly higher than the standard deviations obtained for the

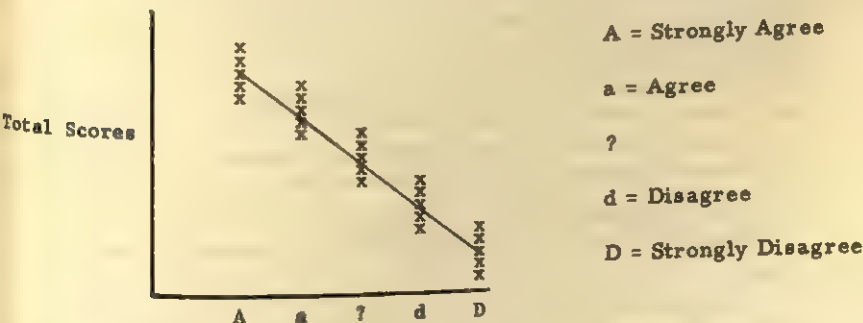


Figure 1. Hypothetical total scores for each response category indicating no difference in variance of total scores.

other response categories' distributions, although the means can still remain unchanged. If, furthermore those motivations do not "counterbalance" each other, then the mean of the "?" distribution of total scores will not be in line with the other means.

This is shown in Figure 2. First, the dispersion of the total scores obtained for those respondents using a "?" is greater than the dispersion of total scores obtained for those responses using any other response category, indicating that the reasons leading respondents to mark the "?" answer are more numerous and more heterogeneous than the reasons which led them to mark any other response category; second, the mean of the "?" distribution does not follow the pattern of means suggested by the other response categories, indicating that the final effect when all the reasons for using the "?" answer are averaged, cannot be considered as equivalent to the one obtained had respondents used the "?" response for the unique purpose of indicating an average position along the agree-disagree continuum.

### *Method*

In order to test which of the above situations was most representative of the way in which respondents interpreted items, data were used which were collected from 300 respondents in connection with a project of a graduate class in Social Attitude Measurement. The class was given the assignment of constructing a series of items both of the Likert type ("Strongly Agree," "Agree," "?", "Disagree," "Strongly Disagree") and polar type<sup>a</sup> ("Strongly Agree with A," "Agree with A," "?", "Agree with B," (B being the polar opposite of A), "Strongly Agree with B"). The items were designed to represent unidimensional

---

<sup>a</sup>Polar choice items are constructed such that two alternative poles of the same dimension are incorporated into one question. Each pole is worded in such a way as to allow any respondent to agree with one of the two poles. The alternatives are chosen such that they are "polarly opposite" in meaning, i.e., agreeing with one alternative logically prohibits a person from agreeing with the other. The format and instructions typically employed are as follows:

Each item consists of two alternatives, A and B, between which you are asked to choose by circling one of the appropriate indicators:

A Statement A is entirely preferred to Statement B as an expression of my opinion.

a Statement A is somewhat preferred to Statement B.

? I cannot choose between A and B.

b Statement B is somewhat preferred to Statement A.

B Statement B is entirely preferred to Statement A as an expression of my opinion.

Example:

A a I feel most people are generally happy.

?

B b I feel more people are generally sad.

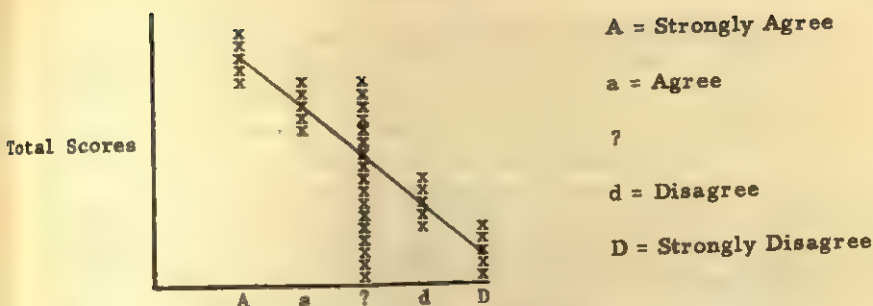


Figure 2. Hypothetical total scores for each response category indicating differences in variance of total scores.

issues in one of the three following areas: childless vs. childful marriage; the generation gap; and the drugs issue. Within each area or team, approximately ten scales were constructed, each one on a different dimension of the same issue. For each dimension within each issue, there was one scale comprising eight polar choice statements, the other scale representing a random ordering of the two derivatives of each polar item, so that each Likert scale contained sixteen items.<sup>4</sup> In total, each team had about 100 respondents, each filling out all questionnaires developed within the team. A criterion-related validity measure and a reliability test were applied to all scales in order to identify the best of them. The 10 best scales were selected, each having a Kuder Richardson formula reliability coefficient of greater than .60 on both Polar and Likert scales, as well as a convergent validity index greater than .20.

Then, for each such scale, total scores were recomputed according to the following scoring system:

Format	Response Category	Coded As	Scored As
Likert	Strongly Agree	A	4
	Agree	a	3
	?	?	2
	Disagree	d	1
	Strongly Disagree	D	
Polar	Strongly Agree with A	A	4
	Agree with A	a	3
	?	?	2
	Agree with B	b	1
	Strongly Agree with B	B	

<sup>4</sup> In this analysis, the criterion for determining who is consistent and who is not is based exclusively on the use of the "?" response category on the polar item. That is, it is inconsistent to agree with Likert A form and disagree with Likert B form, and at the same time have selected the "?" on the polar item (this pattern should have followed from agreeing with the A side of the polar choice question). Conversely, it is consistent to disagree or agree with both Likert items and mark a "?" for the corresponding polar statement.

Thus, the contributions of the question mark response category to the total scores were ignored. This was necessary since scoring responses the traditional way would have implied that the question mark response category is an indicator of an average position.

As a result of this modification, the total scores of those respondents who used at least one "?" answer were based on less than the total number of items of the scale. Thus, to allow the comparisons of scores based on a different number of contributing items, *average* rather than *total* scores were used. They were computed according to the following formula:

$$\text{Average total score} = \frac{\text{Total score computed on a 4-point scale}}{\text{Total number of items—Number of items for which the respondent used the "?" response category}}$$

Finally, before comparing the variances of average total scores obtained for each response category of each item, it was necessary to remove the contribution of the responses for the item under consideration from the average total score. Otherwise, its inclusion would have resulted in double counting and the data would have been "contaminated." Thus for each item, average total scores were recomputed as follows:

Response to the item under consideration		
Polar Format	Likert Format	
A	A	New Average total score = $\frac{\text{Total score} - 4}{\text{No. of items} - 1}$
a	a	New Average total score = $\frac{\text{Total score} - 3}{\text{No. of items} - 1}$
?	?	Average total score (unchanged) = $\frac{\text{Total score} - 0}{\text{No. of items} - 0}$
b	d	New Average total score = $\frac{\text{Total score} - 2}{\text{No. of items} - 1}$
B	D	New Average total score = $\frac{\text{Total score} - 1}{\text{No. of items} - 1}$

After all these modifications, the variance of the new average total scores "?" response category was compared to a weighted average of the variances of the new average total scores obtained for the other

response categories and the corresponding  $F$  ratio was tested for statistical significance.

### *Results*

As can be seen in Table 1 the average total score variance of the "?" category is significantly higher (at the .05 level) than the average weighted variance computed for all other categories in only four cases out of 240, and significantly smaller in 11 cases. Therefore, it can be concluded that the variability of the average total scores of respondents selecting the "?" response category is not statistically different from the variability of the other categories. The average variability for each response category shown in Figure 3 confirms this result.

Furthermore, Figure 3 shows that the means of average total scores obtained from each category across both types of formats are virtually on a straight line, and that the mean corresponding to the "?" response category holds an intermediate position with respect to the other means, indicating that those choosing the "?" category are those who are in the middle of the attitude continuum as measured by the average total scores. Therefore, the assumption underlying the scoring of the "?" response category made by attitude researchers when coding their items is supported by empirical evidence across ten different scales in both polar and Likert formats.

To test the scope and degree of stability of this conclusion, it was decided to observe the behavior of the responses obtained in a context where the potential uses of the "?" response category are increased and its meaning further delineated. Such a context is provided by the simultaneous comparison of each polar choice question and its two Likert derivatives. Thereby the pattern of responses of individuals across essentially three similar questions can be followed.

#### *Analysis of Question Mark Response across Polar and Likert Formats*

Table 2 illustrates the distribution of responses on the two Likert questions, given the selection of a "?" on the polar choice format. As can be seen, respondents choosing the "?" response category on a polar choice question have a variety of reasons for doing so.

The largest individual cell in the table is the one predicted from the hypothesis that a person choosing the "?" for the polar choice question will also choose a "?" in the two Likert derivatives of the same item. Eighteen percent of the respondents fell in this category. This response pattern represents those who may be called the "truly un-

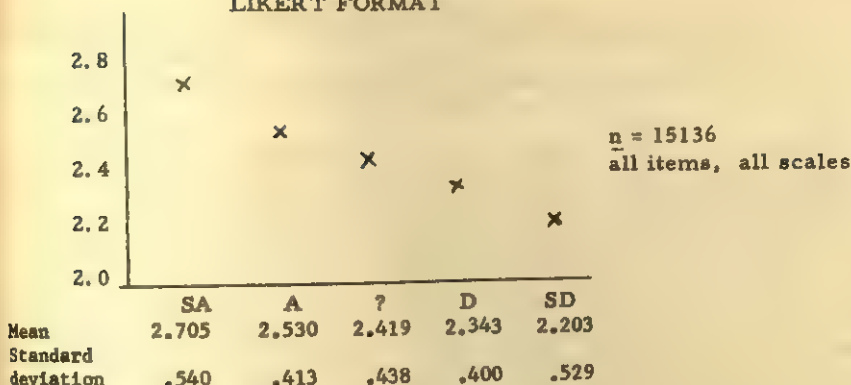


TABLE 1  
Comparison of Variances of Average Total Scores across Scales  
F-Ratios

Scale No.	1	2	3	4	5	6	7	8	9	10
Item No.	1	2	3	4	5	6	7	8	9	10
Polar Format	1	1.532	.619	.657	2.856*	1.260	.743	1.500	3.058	1.481
	2	1.756	1.853	.916	1.583	.547	.387	1.300	.761	.647
	3	1.680	.940	.839	1.407	1.786	.838	.747	1.374	.654
	4	1.411	1.946	1.003	.716	1.579	1.510	.673	.599	2.169*
	5	1.543	1.226	1.326	.113*	1.287	1.472	1.264	.593	1.029
	6	1.631	.698	1.308	.816	1.210	2.224	.819	1.214	.759
	7	2.131*	.860	1.101	1.507	1.824	2.004	2.403	1.945	.763
	8	1.056	1.227	.924	1.101	1.448	2.054	.681	.887	2.689
Likert Format	1	1.791	.615	.633	1.542	.186*	.776	.790	.688	1.114
	2	1.591	.320*	1.211	.972	.715	1.527	1.230	1.198	1.046
	3	.390*	.531	1.164	.968	1.022	.414	.773	.214*	1.108
	4	2.371	.738	.813	—	.753	.981	.432	.923	1.186
	5	1.119	1.146	1.076	.424	1.130	1.097	.118*	.972	.832
	6	.542	.511	.275*	1.522	1.788	1.199	2.041	.809	1.133
	7	1.558	.727	1.140	1.837	.445	.722	.740	.436	1.261
	8	.748	.839	.934	.312*	1.218	.518	.176*	.604	1.324
	9	.498	.726	.777	1.938	.693	.921	1.083	.804	.006
	10	1.181	1.250	1.173	1.250	1.576	.637	.431	1.310	1.748
	11	.808	.436	.755	1.114	.675	.980	.791	.702	1.013
	12	1.698	.586	.769	1.193	.782	—	.934	.877	.642
	13	1.117	.739	1.752	1.158	1.283	1.438	.972	.786	1.054
	14	.467	.573	1.338	1.077	2.233	1.236	.623	.510	1.607
	15	1.212	.631	1.061	1.288	2.023	.402	1.291	.590	.351*
	16	1.731	1.493	.774	1.243	.916	.758	1.077	.219*	2.981*

\* =  $p \leq .05$  or  $\geq .95$ .

## LIKERT FORMAT



## POLAR FORMAT

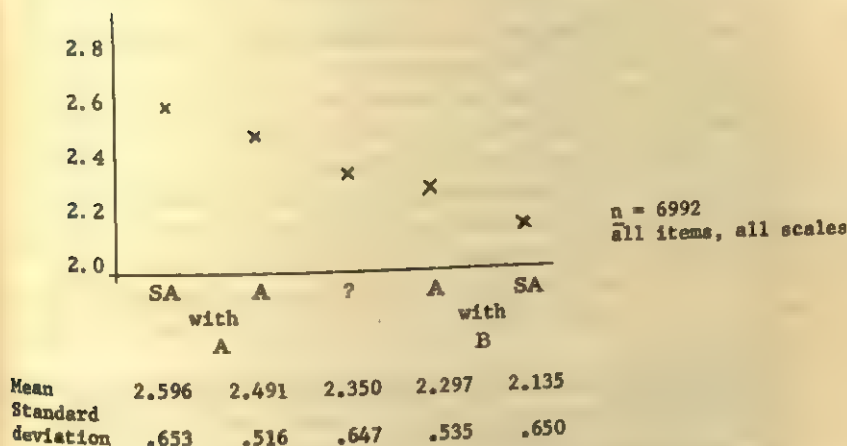


Figure 3. Means of average total scores obtained for each response category.

decided." The label "truly undecided" may be a misnomer, however, since this 18% may not actually represent an equal and infrequent use of the "?" by a majority of respondents, but may be due to a few persons consistently employing a "?" across different question contents. Such a pattern of responding has received wide attention in the psychometric literature (e.g., Cronbach, 1946; Bentler, Jackson, and Messick, 1971) under the general name of response styles.

It is unlikely that a person will consistently respond with a "?" across scales dealing with relatively different topics. Accordingly, as a rough test of the existence of evidence that some persons may be "?"

TABLE 2

*Distribution of Responses across Both Likert Scales, Given the Selection of a "?" on the Corresponding Polar Item*

Count Tot. Pct.	LIKERT B					Row Total
	Strongly Agree 1	Agree 2	"?" 3	Disagree 4	Strongly Disagree 5	
Strongly Agree	8	13	24	17	27	89
1	.6	1.0	1.9	1.3	2.1	7.0
Agree	8	47	71	74	27	227
2	.6	3.7	5.6	5.8	2.1	17.9
Likert A	8	38	225	78	29	378
3	.6	3.0	17.5	6.1	2.3	29.8
Disagree	13	89	83	157	36	378
4	1.0	7.0	6.5	12.4	2.8	29.8
Strongly Disagree	20	26	32	35	85	198
5	1.6	2.0	2.5	2.8	6.7	15.6
Column Total	57	213	435	361	204	1270
	4.5	16.8	34.3	28.4	16.1	100.0

response stylers, a simple frequency count of the number of times someone chooses a "?" was performed.

From this frequency count, it was found that the 18% of the responses found in this cell was not 18% of the sample as would be expected if the hypothesis of no "?" response styles were true. Instead, 9% of the respondents in our sample (or 22% of the number of people comprising the initial 18%) accounted for 55% of the triple "?" thereby lending credibility to the hypothesis of the potential presence of a "?" response style. Additional support for this hypothesis can be found as well in the psychometric literature (cf. Rosenberg, Ixand, and Hollander, 1955).

All other responses appearing in Table 2 can be categorized as those who are rationally consistent<sup>a</sup> on both Likert questions, or those, who, if they understood the question, were inconsistent, given that they accurately selected the "?" on the polar format.

Persons who are *consistent* are the type of people who, when choosing "?" on the polar choice will agree with one Likert question as well as agreeing with the other, the "polar opposite." Six percent of total responses fell in this category. Another possible response pattern for "consistent respondents" is disagreeing with both Likert questions. Almost one-fourth of the responses fall into this category. This dis-

X X X X X X X X X X X X X X X X  
<sup>a</sup> For example, the "happy/sad" polar item would have been translated into the two following Likert format statements:

I feel most people are generally happy  
 I feel most people are generally sad

A a ? d D  
 A a ? d D

crepancy found between these two percentages may be due to the specific format under which the polar choice method is developed. This assumption seems supported by the fact that the number of disagreements with each Likert item was higher than the number of agreements (44.5% versus 21.3% and 45.4% versus 24.9%) as can be seen in the marginals in Table 2. Were the polar choice format constructed with negative polar opposites instead of positive opposites (i.e., respondents would be asked to disagree with one of the two polar statements) the reverse trend would probably occur, i.e., more people would be consistent in agreeing rather than consistent in disagreeing.

All persons falling in a "?" category on either Likert item can be considered as *rational* if their reason for selecting a "?" on one Likert scale is stronger than their reason for selecting another response category on the other Likert item. Otherwise, they can be considered as *irrational* responders. Since "Strongly Agree" represents a firmer commitment than "Agree" it seems that those persons who strongly agreed with either Likert should not have used a "?" in the first place (unless their reasons for selecting the "?" on the other Likert are very strong) and probably fall into the category of "irrational responders." The same argumentation can be used to tentatively classify those who used a "a" on one Likert and a "?" on the other as "rational."

The final category is that of the *inconsistent* respondents who are those who agree with one Likert and disagree with the other: they should not have selected the "?" on the polar item in the first place.

Table 3 presents a more concise summary of the basic types of responders. As can be seen, just under 40% select the question mark

TABLE 3.  
Basic Types of Responses when a "?" is Selected on a Polar Item

Item Response on Likert 1	Item Response on Likert 2	Respondent Segment	Percentage of all responses
?	?	truly undecided and/or response stylers	17.7%
?	a or d	rational responders	21.2%
a or d	?		5.9%
A or a	a or A	consistent responders	24.7%
D or d	D or d		7.3%
A or D	?	irrational responders	11.3%
A or a	d or D		11.6%
D or d	A or A	inconsistent responders	30.2%

for the reason most researchers assume, and the remaining sixty per cent are split between being consistent and being what is called here, irrational, or inconsistent.

### *Analysis of Responses on Both Likert Questions Only*

Table 4 presents similar results for the two Likert questions. As can be seen, the most popular response again is the "?" on the second Likert, given its selection on the first. Although this is to be expected, as mentioned above some of these responses may be due to the presence of a "?" response style. The next most frequent categories are those on either side of the question mark, and finally only a small minority select the extreme responses. This, again, is generally consistent with the scoring pattern employed by most researchers.

### *Discussion*

The foregoing sections show how some understanding of the meaning of the "?" response category can be achieved from an analysis of responses across different item formats as well as items' average total scores and response category variability. If further insight is desired, it is suggested that researchers extend for example the study by Goldberg (1971) in allowing respondents to state their reasons for selecting the "?" (e.g., they would allow respondents to indicate whether they are indifferent; have mixed feelings; the question is ambiguous; the responses don't fit the question, etc., and in addition in the polar format that they disagree or agree with both poles). Also, it would be of

TABLE 4  
*Analysis of "?" Responses across Likert Questions*

Responses on LIKERT A from "?" on LIKERT B		Responses on LIKERT B from "?" on LIKERT A	
A	50 5.9%	A	90 8.8%
a	190 22.5%	a	250 24.8%
?	276 44.5%	?	376 37.3%
d	160 18.9%	d	208 20.6%
D	69 8.2%	D	82 8.5%
Column Totals	845 100.0%		1006 100.0%

Note—Numbers appearing in each cell represent the count and percent of total (column) responses.



interest to record respondents' reactions to being able, and not being able to select a middle category on a set of items (i.e., scale items would be presented under both five and four-point formats). Responses to items in both formats then would be compared as well as respondents' reactions to the entire scale. It would also be interesting to analyze the frequency of the use of the "?" when the number of response categories are systematically varied.

Finally, the investigation of the existing relationship between perceived knowledge or familiarity and interest in the area under analysis and frequency of use of the "?" would provide additional insight into the respondents' use of the question mark response category.

The results reported in this paper lead to the conclusion that the variability of meanings of the "?" response is not greater than the variability of the other response categories. Even though the current practice of including a "?", and treating it as the middle category seems supported, it does not mean that "?" is *only* an indicator of an average position as most investigators assume. For example, some evidence exists for the presence of a "?" response style and of ambivalent or indifferent responses.

In conclusion, this study suggests that a researcher would be advised to check the variability of his "?" responses against the other response categories. This would indicate the range of meanings attached to the "?", as compared to the other response categories. Next, by looking at the mean of the "?" response distribution, he would know what appropriate score must be given to these responses. In other words, the researcher should be more careful in specifying the meaning of the category rather than just naively assuming that it indicates an average position.

Despite its wide use in attitude scales, the question mark, or any other response category for that matter, has received far less theoretical consideration than needed if researchers want to validly attribute meaning of responses to items in attitude scales. It is hoped that this study represents a step in the analysis of the "?" responses which will stimulate other researchers to pursue in this direction, and more generally will sensitize users of attitude scales to an important methodological issue, that of determining empirically the meaning respondents attribute to their responses to an item.

## REFERENCES

- Bentler, P. M., Jackson, D. N., and Messick, S. The identification of content and style: A two-dimensional interpretation of acquiescence. *Psychological Bulletin*, 1971, 76, 86-204.

- Brown, R. *Social psychology*. New York: The Free Press, 1965.
- Cook, S. and Selltiz, C. A multiple indicator approach to attitude measurement. In *Attitude Measurement*, G. F. Summers (Eds), Rand McNally, Chicago, 1970.
- Cronbach, L. J. Response sets and test validity. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1946, 6, 475-494.
- Diab, L. N. Some limitations of existing scales in the measurement of social attitudes. *Psychological Reports*, 1965, 17, 427-430.
- Edwards, A. L. A critique of "neutral" items in attitude scales constructed by the method of equal appearing intervals. *Psychological Review*, 1946, 53, 159-169.
- Goldberg, G. Response format in attitude scales. Unpublished manuscript, Northwestern University, 1971.
- Kaplan, K. J. On the ambivalence-indifference problem in attitude theory: A suggested modification of the semantic differential technique. *Psychological Bulletin*, 1972, 77, 361-372.
- Likert, R. A technique for the measurement of attitudes. *Archives of Psychology*, No. 140 (1932).
- McDonald, R. A unified treatment of the weighting problem. *Psychometrika*, 1968, 33, 351-382.
- Oppenheim, A. N. *Questionnaire design and attitude measurement*. New York: Basic Books, 1966.
- Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. *The measurement of meaning*. Urbana, Illinois: University of Illinois Press, 1957.
- Stanley, J. C. and Wang, M. D. *Differential weighting: A survey of methods and empirical studies*. New York: College Entrance Examination Board, 1968.
- Rosenberg, N., Ixand, C., and Hollander, E. Middle response category: reliability and relationship to personality and intelligence variables. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1955, 15, 281-290.
- Summers, G. F. (Ed.) *Attitude measurement*. Chicago: Rand McNally, 1970.
- Wang, M. D. and Stanley, J. C. Differential weighting: A review of methods and empirical studies. *Review of Educational Research*, 1970, 40, 663-705.
- Webb, E. J., Campbell, D. T., Schwartz, R. D., and Sechrest, L. *Unobtrusive Measures: Non-reactive Research in the Social Sciences*. Chicago: Rand McNally, 1966.
- Worthy, M. Note on scoring midpoint responses in extreme response-style scores. *Psychological Reports*, 1969, 24, 189-190.

## SECTION SELECTION IN MULTI-SECTION COURSES: IMPLICATIONS FOR THE VALIDATION AND USE OF TEACHER RATING FORMS

LES LEVENTHAL, PHILIP C. ABRAMI, RAYMOND P. PERRY  
AND LAWRENCE J. BREEN<sup>1,2</sup>

University of Manitoba

Researchers know little about determiners of section selection in multi-section college courses. Studies on teacher evaluation and on the validity of teacher rating forms have often assumed section to section equivalence of students assigned by customary registration procedures. To investigate the section selection process, a questionnaire containing items on personal history, reasons for section selection, and sources of information about the instructor was administered to 1,188 undergraduate students in multi-section first year and advanced psychology courses. Major findings were: (1) students significantly differed across sections on biographical variables and on section selection reasons, (2) time at which class was scheduled (classtime) and teacher's reputation were the primary reasons for section choice, (3) teacher's reputation was less important than classtime for first year students, but comparable to classtime for advanced students, and (4) reports from other students and published ratings were, respectively, the first and second most frequent source of instructor reputation information.

CAMPBELL and Stanley (1963) have noted that many educational researchers have analyzed studies that use administrative assignment

---

<sup>1</sup> This research was supported in part by grants from the University of Manitoba Research Board and the University of Manitoba Group for Advance in Innovative Instruction. The authors express their gratitude to W. J. McKeachie who reviewed an earlier version of this paper which was presented at the meeting of the Canadian Psychological Association, Windsor, 1974.

<sup>2</sup> Requests for reprints should be addressed to Les Leventhal, Department of Psychology, University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2.

of students to classes as though random assignment had been employed, and therefore often have reached incorrect conclusions about treatment effects. Research into section selection and the equivalence of students in different sections is crucial to the literature on teaching effectiveness and on the validity of teacher rating forms (e.g., Costin, Greenough, and Menges, 1971). Researchers who have tried to validate a teacher rating form (TRF) frequently have assumed that differences among teachers on TRF ratings are due to teaching effectiveness rather than to initial student differences. They typically have studied multiple sections (taught by different teachers) of the same course and have computed the correlation between section means on a TRF and section means on validity criteria, such as a common final examination. With few exceptions (Sullivan and Skanes, 1974), students have *not* been randomly assigned to sections.

Leventhal (in press) has argued that performance differences among sections on validity criteria may be due to factors other than teacher ability such as student ability and motivation. Furthermore, Leventhal has suggested that even when teachers have no effect on validity criteria, a correlation between TRF means and validity criteria means, which may nevertheless occur because of the lack of randomization, may mistakenly be used as evidence for TRF validity. In the present investigation an attempt was made (a) to examine the section selection process to determine whether the process approximates random assignment and (b) to assess whether the process for advanced undergraduates differs from that for first year undergraduates.

### *Method*

#### *Subjects and Setting*

The subjects for this study were 940 students from 13 Introductory Psychology sections and 248 advanced undergraduate students from 6 sections of Social Psychology. During the academic year prior to this study, the University of Manitoba Student Union (UMSU) constructed, administered, and analyzed a TRF for all departments in the Faculty of Arts, which includes Psychology. The results of these ratings were published and made widely available at the time students registered for the courses investigated in this study.

#### *Materials and Procedure*

A 22-item questionnaire was constructed in which the first eight items related to the following student demographic characteristics: (1)



education, (2) age, (3) sex, (4) family income, (5) hometown population size, (6) college grade-point average (GPA), (7) high school GPA, and (8) an indication of whether or not the course was required. Items 9 to 16 assessed specific reasons for section choice, which included: (9) identifiability and importance of section choice reason (clarity of reason), (10) classtime (scheduled time), (11) desire to be with friends, (12) classroom location, (13) physical features of classroom location, (14) nature of assigned reading materials, (15) teaching ability and/or reputation as a teacher, and (16) any reason other than those listed. Response alternatives to section selection items ranged on a 4-point scale from "selection based entirely upon" the reason (one) to "selection not at all based upon" the reason (four). An "I don't remember" alternative was also provided. The remaining items, 17 to 22, related to the following kinds of knowledge which respondents had about their instructor at the time of registration: (17) awareness by student of his/her instructor's published TRF evaluation, (18) information regarding whether or not the instructor had been evaluated by UMSU, (19) amount of pre-enrollment information about instructor's teaching ability and/or reputation, (20) source of information, (21) pre-enrollment conclusion about instructor, and (22) accuracy of pre-enrollment information. The questionnaire was administered three to four weeks into the regular session term.

## *Results*

### *Reasons for Section Selection*

A mean and standard deviation for each course on items 9 to 15 were computed (see Table 1). The low means for items 9 show that students in both courses generally maintained that they had a clear reason for their section choice. Of the specifically named reasons (10 to 15), time of scheduling (classtime) and teacher's reputation were the primary reasons for both courses. Although time of scheduling was more important than reputation for Introductory Psychology students, time was comparable to reputation in importance for Social Psychology students. Furthermore, the rank ordering of importance of the set of reasons was similar for both courses. For each of the specifically named reasons, a one-way analysis of variance (ANOVA) was computed. Results indicated that Social Psychology students declared that they had more clearly identifiable reasons for section selection than Introductory students ( $F = 13.96$ ,  $df = 1/1168$ ,  $p < .001$ ). None of the specific reasons significantly differed in importance for the two courses except reputation of the professor. Social Psychology



TABLE 1  
*Introductory Psychology and Social Psychology Students  
 Compared on Section Selection Reasons*

Item	<i>n</i> <sup>a</sup>		<i>M</i>		<i>SD</i>		<i>F</i>	<i>p</i>
	Intro. Psych.	Social Psych.	Intro. Psych.	Social Psych.	Intro. Psych.	Social Psych.		
9. Clear reason	940 (921)	248 (248)	2.27	1.96	1.17	1.06	13.96	<.001
10. Time	940 (921)	248 (245)	2.89	2.75	1.09	1.00	3.21	.07
11. Friends	940 (923)	248 (247)	3.83	3.89	0.53	0.41	3.34	.06
12. Location	940 (925)	248 (247)	3.74	3.78	0.58	0.46	1.10	.29
13. Room								
Features	940 (915)	248 (246)	3.97	4.00	0.22	0.06	2.97	.08
14. Readings	940 (920)	248 (246)	3.77	3.72	0.53	0.53	1.34	.24
15. Ability/ Reputation	940 (928)	248 (245)	3.33	2.84	1.00	1.04	44.34	<.001

<sup>a</sup> Numbers in parentheses indicate number of responses analyzed after defective responses and "I don't remember" responses were dropped.

students reported reputation to be more important to section selection than did Introductory students ( $F = 43.3$ ,  $df = 1/1172$ ,  $p < .001$ ).

To obtain a simple overall picture of the frequency with which students in each course cited specific reasons to be of dominating importance, for each course the percentage was computed of students who had based their section selection (a) *mostly* or *entirely*, or (b) *not at all* on that reason (Table 2). The data again indicate that two major reasons, classtime and reputation, as well as a collection of minor reasons apparently influence students' choice of sections. But even the most potent reason, classtime for Introductory Psychology students or reputation for Social Psychology students, mostly or entirely in-

TABLE 2  
*Reasons Reported by Students for Section Selection*

	% basing decision mostly or entirely on reason		% basing decision not at all on, or failing to recall, reason	
	Intro Psych.	Social Psych.	Intro Psych.	Social Psych.
Time of Class				
Teacher's ability/reputation	34.3	34.4	40.2	25.9
Readings	20.9	38.5	63.5	38.5
Location	3.8	3.2	81.5	76.1
Friends	4.8	1.2	80.1	79.8
Room Features	4.1	2.0	88.5	92.3
	0.6	0.0	98.0	99.6

*Note.*—Since students responding "somewhat" to the importance of a reason were excluded, the percentages on a row for a course do not total 100%.

fluences the decisions of fewer than 40% of the students, and a comparable percentage of students ignores the reason altogether. These data, therefore, reveal that although reasons greatly differ in importance, no single reason mostly or entirely determines the section selection for the majority of students.

### *Approximation to Random Assignment*

To determine whether the section selection process for a course approximates the random assignment of students to sections, analyses were made of section selection and biographical items. Introductory Psychology sections were analyzed separately from Social Psychology sections. One-way analyses of variance (ANOVAs) were computed across sections for each question (see Table 3) even though certain questions have alternatives that do not meet the formal interval data assumptions of the ANOVA (Burke, 1963). Taken as a whole, the data

TABLE 3

*Probabilities Associated with F Tests for Analyses of Variance Computed for Each Question across Sections*

Item	Introduction to Psychology <i>n</i> = 940	Social Psychology <i>n</i> = 248
Demographic		
1. Education	<.001	.053
2. Age	<.001	.324
3. Sex	<.001	.184
4. Family Income	.681	.553
5. Hometown population	.001	.200
6. College GPA	.616	.117
7. High School GPA	.049	.683
8. Course required	<.001	.194
Section selection		
9. Clear reason	.013	.004
10. Time	.001	.001
11. Friends	.641	.221
12. Location	.001	.029
13. Room features	.388	.550
14. Readings	.317	.022
15. Ability/reputation	<.001	<.001
16. Another reason	<.001	.448
Information on instructor		
17. Looked up TRF ratings	<.001	.080
18. Instructor previously rated	<.001	.011
19. Amount pre-enrollment information	<.001	<.001
20. Source of information	.060	.027
21. Pre-enrollment conclusion	<.001	<.001
22. Accuracy of information	<.001	<.001

in Table 3 show that the sections within a course significantly differed from each other on both biographical dimensions and section selection reasons. This outcome is especially clear in the analysis of the Social Psychology course. Though only six Social Psychology sections were tested, five out of the seven items that relate to reasons given by students for choosing their sections (items 9 to 15) showed significant differences across sections beyond the .05 level, and three out of seven were significant beyond the .01 level. Furthermore, a high degree of agreement between the two courses was shown on which of the seven questions yielded significant section differences.

### *Sources of Students' Information*

Combining all students from both courses, students revealing that they had had pre-enrollment information about their prospective instructor's teaching ability or reputation were asked to indicate the source of this information and their conclusion about the prospective instructor. Cross-tabulation tables were prepared and translated into the form of a chart that related sources of pre-enrollment information and pre-enrollment conclusions about prospective professors (see Figure 1). These data suggest that regardless of pre-enrollment conclusion about the instructor, a student will *retrospectively* recall the following as useful sources of pre-enrollment information (most frequently recalled source first): comments from other students, UMSU booklet, other sources, pre-enrollment audition, and previous experience with the instructor. In addition, other students were seen as a *far* more important source of information than were the remaining sources. Finally, all sources of information generally appeared to be about equally useful sources of favorable and unfavorable information with the exception of "other students": the more favorable a student's pre-enrollment conclusion, the more likely the student would maintain that other students had been the source of useful information.

### *Accuracy of Information*

Analysis of the accuracy of pre-enrollment information item showed that 31.0% stated that information was "totally accurate," that 46.5% declared "mostly accurate," that 20.0% reported "somewhat accurate," and that 2.6% cited "completely wrong." To determine where students in each of these designated response categories obtained their information, cross-tabulation tables were prepared and then cast in the form of a chart that related post-enrollment judgment about accuracy of pre-enrollment information about their instructors (see Fig-

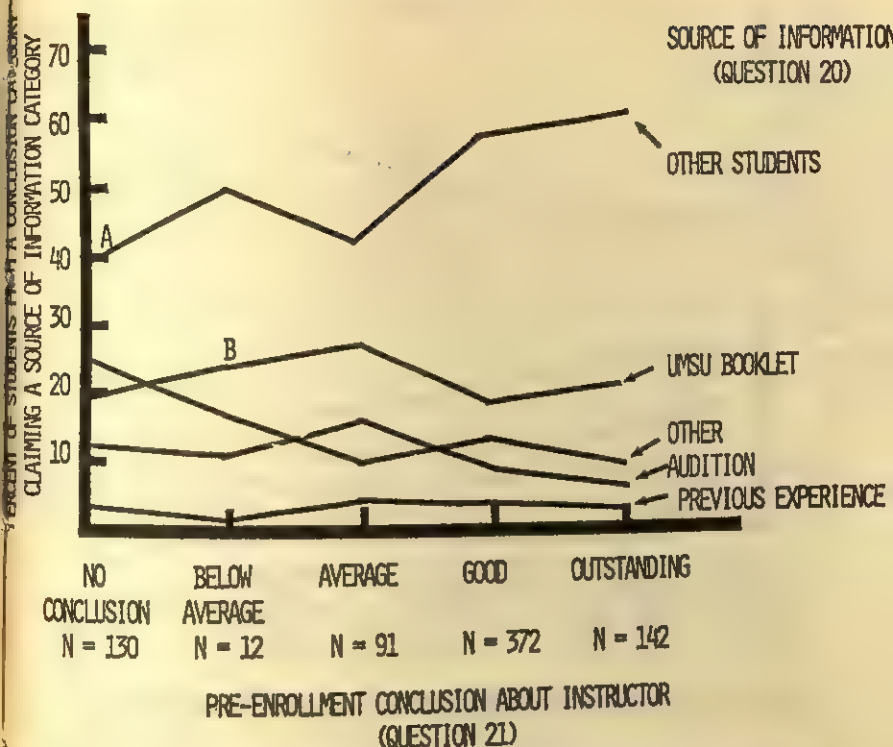


Figure 1. Sources of information about instructors claimed by students reaching certain pre-enrollment conclusions about their instructor.

ure 2). These data suggest that all sources of information were judged accurate by some students and inaccurate by others. In addition, the more a student judged his pre-enrollment information to be accurate, the *more* likely he was to credit other students as an informative source and the *less* likely he was to credit published UMSU ratings.

Analysis of responses to item 9 showed that about 80% of all students reported that their reasons for selecting their section were somewhat, mostly, or clearly identifiable. An analysis of item 16 showed that about 72% declared that their section selection was not based, or only somewhat based, upon a different reason from those reasons listed in the questionnaire. In short, a large majority of students had a specific reason for choosing their section—a reason that was listed in the questionnaire. The final portion of item 16, which consisted of a fill-in blank, requested students to describe any other reasons than those already listed for choosing their section. Of all students tested,

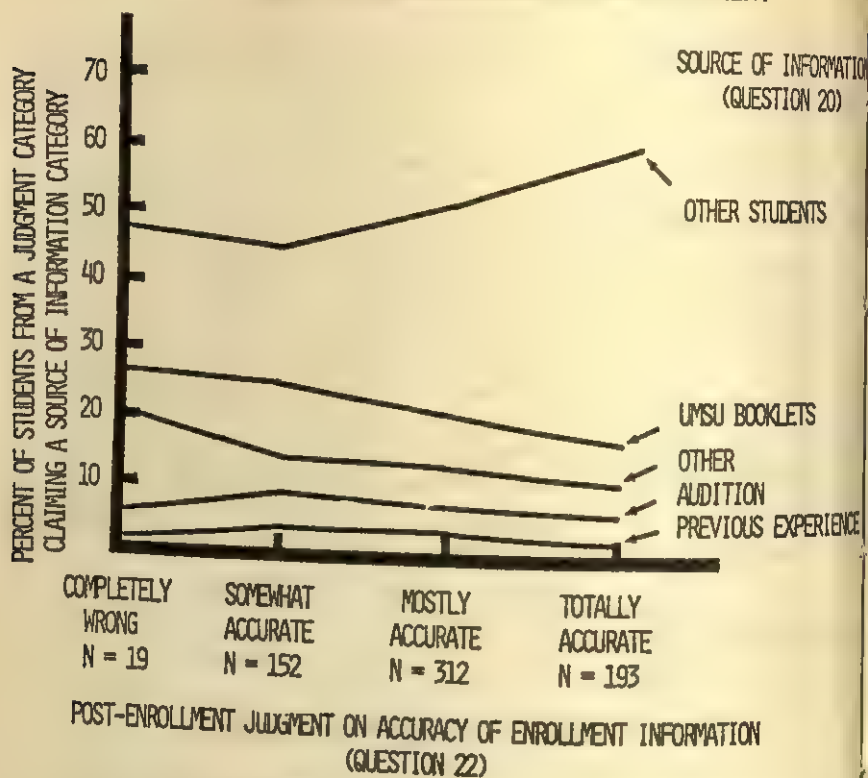


Figure 2. Sources of information about instructors claimed by students reaching certain post-enrollment judgments on the accuracy of that information.

72% ignored the fill-in blank, 12% reported that they had no choice other than to select the section, 10% apparently did not understand item 16, and 6% listed a reason different from those other reasons in the questionnaire.

A correlation coefficient computed over section selection items between mean importance of item for all sections combined and probability of random section to section variation in section means on item provides a conservative estimate of the degree to which responses to the questionnaire are controlled by variables related to actual section selection. Such a correlation was computed over items 10 to 15 between the mean importance of item for all sections combined, and the probability of obtaining an  $F$  ratio at least as large as the one found for section to section variation in sections means. The correlation coefficient was .64 for Introductory Psychology, .62 for Social Psychology, and .73 for both courses combined.



### *Discussion*

#### *Section Selection Process*

The results of this study show that classtime (scheduled time of class) and teacher's reputation are the first and second most important variables controlling section choice. Although reputation lags in importance considerably behind classtime for freshman students, reputation is comparable to classtime for advanced undergraduates. The instructor reputation data have implications for student evaluations of teachers. Perry, Niemi, and Jones (1974) have found that students more favorably rate highly reputed teachers than they rate poor ones. The present data indicate that many students admitted seeking reputation information prior to registration—a tendency that was stronger with advanced undergraduates than with less advanced students. In addition, since the present data reveal that students significantly varied across sections in their statements of the importance of a teacher's reputation, it appears that this declared reason was, in fact, relevant to section choice. Thus, one may infer that teachers tend to be locked into their reputations and that this tendency is stronger in advanced courses than in introductory ones. More importantly, the present data demonstrate that students significantly varied from section to section along a number of biographical dimensions and according to their reasons for section choice. In short, the assumption that the section selection process is sufficiently complex to approximate random assignment was not supported by this study.

Furthermore, it would appear that many of these factors may influence either the TRF ratings by students or their performance on TRF validity criteria. For this reason, failure to randomize students may produce inaccurate estimates of teachers' impact on student performance on completing TRF's and on TRF validity criteria. The present findings that students vary from section to section on more than just one dimension and that no section selection reason mostly or entirely influences the decisions of more than 40 per cent of the students implies that the section selection process is controlled by many important variables rather than by a single dominating one. Hence, non-randomized TRF validation studies that use statistical control techniques (e.g., part and partial correlation) to control for initial student differences among sections must control many student variables. Typically, these studies have controlled only student ability (e.g., Elliott, 1950).

#### *Sources of Students' Information*

The present data were collected at a university where published TRF results were available to students during course registration. If

student differences across sections are primarily due to selection of highly reputed teachers, and if interest in teacher reputations is due to the availability of published ratings, then the present data may be representative only of institutions that publish TRF evaluations. On the other hand, if students acquire information about teachers from other sources—e.g., other students—then the present data may have wider generality than they would otherwise have. The present investigation shows that other students were the most frequent source of information about instructors; hence, the section selection processes identified in this study may also occur at institutions where published TRF results are unavailable.

Generally, all sources of information except "other students," were remembered to be equally useful sources of favorable and unfavorable knowledge for decision-making. Hence, there is no strong evidence that the availability of published TRF results would greatly change the balance of favorable and unfavorable information available to students.

#### *Cautions in Interpreting Results: Implications for Valid Use of a TRF*

The checks on the adequacy of the questionnaire indicate that the questionnaire did not omit any important reasons for section choice and, more importantly, that responses to the section choice items reflected the actual selection process. Nevertheless, the present data should be interpreted with caution so that their validity may not be misrepresented. First, these data are correlational; causal interpretations must be made with care. Second, all data are *retrospective* reports vulnerable to distortions and colorations. For example, a source remembered as providing useful information may not have influenced section selection. Third, many of the present findings achieve practical significance, not because the variables studied were shown to be of dominating importance to section selection, but because they are sufficiently potent to provide re-interpretations of existing research which typically involves large groups and uses powerful statistical techniques. For example, only 21% of Introductory Psychology students studied maintained that they had based their section choice mostly or entirely on a teacher's reputation (see Table 2). Nevertheless, such a percentage may produce significant section to section differences in student characteristics associated with interest in teacher's reputation when the differences are computed in a study employing nonrandom assignment to sections of more than 20 students. For the various reasons cited, the appropriate use of the results obtained from a TRF may be difficult to ensure in specific college and university settings.

## REFERENCES

- Burke, C. J. Measurement scales and statistical models. In M. Marx (Ed.), *Theories in contemporary psychology*. New York: MacMillan Company, 1963.
- Campbell, D. T., and Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963.
- Costin, F., Greenough, W. T. and Menges, R. J. Student ratings of college teaching: Reliability, validity, and usefulness. *Review of Educational Research*, 1971, 41, 511-535.
- Elliott, D. N. Characteristics and relationships of various criteria of college and university teaching. *Purdue University Studies in Higher Education*, 1950, 70, 5-61.
- Leventhal, L. Teacher rating forms: Critique and reformulation of previous validation designs. *Canadian Psychological Review*, in press.
- Perry, R. P., Niemi, R. R., and Jones, K. Effect of prior teaching evaluations and lecture presentation on ratings of teaching performance. *Journal of Educational Psychology*, 1974, 66, 851-856.
- Sullivan, A. M., and Skanes, G. R. Validity of student evaluation of teaching and the characteristics of successful instructors. *Journal of Educational Psychology*, 1974, 66, 584-590.



## ESTEEM CONSTRUCT GENERALITY AND ACADEMIC PERFORMANCE

C. KENNETH SIMPSON AND DAVID BOYLE<sup>1,2</sup>

Cleveland State University

Measures of global, specific, and task-specific self-esteem were administered to 78 male and 81 female college students and related to predicted and actual performance on a midterm examination. Significant correlations were found between global and specific measures and between specific and task measures, but not between global and task measures. The relationship between the esteem measures and actual performance was strongest for the task measures, next strongest for the specific measures, and nonsignificant for the global measures. Specific measures were also significantly related to predicted performance, but global measures were not. The findings were discussed in terms of four criticisms of global measures, and it was suggested that more specific self-esteem measures be developed.

THE three major self-esteem constructs identified in the literature thus far can be seen as representing different levels of generality in an hierarchy of esteem constructs. *Global self-esteem*, usually defined as an individual's evaluation of his overall worth as a person, is assumed to be the weighted function of esteem in more specific areas. *Specific self-esteem* refers to evaluations either made in certain life situations (social interaction, male-female relations, education, work) or based on particular aspects of the individual (physique, intelligence, personality, interpersonal competence). But each of these sources is still rather broad, since it includes a multitude of different behaviors and

<sup>1</sup> Requests for reprints should be sent to C. Kenneth Simpson, Department of Psychology, Cleveland State University 44115.

<sup>2</sup> The authors wish to thank Tom Incorvia and David Droll for their assistance with the data analysis and Sam Lane and John Wilson for their assistance in the preparation of this article.



situations. *Task-specific* or *situational self-esteem* refers to evaluations of more restricted sets of behaviors in specific situations. This constraint can be conceptualized in one way as the expectations by the individual of his performance in a task-specific situation.

In two experimental studies the relationship between both global and specific self-esteem measures and behavior has been compared. Schrauger (1972) employed a global measure and a measure of task-specific self-esteem to study the behavior of high and low esteem subjects performing a concept formation task alone or in the presence of an audience. Both measures were related significantly to subjects' perceptions of performance and confidence ratings, but only the task-specific measure was related to actual performance. Morrison, Thomas, and Weaver (1973) found that both the global measure and a specific measure (school esteem) from Coopersmith's Self-Esteem Inventory (SEI), but not the global Social Self-Esteem measure of Ziller, Hagey, Smith, and Long (1969), were significantly correlated with both predicted and actual performance on a midterm examination. However, when predictions of performance made after the test were adjusted through using actual grades as the covariate, only the global SEI measure produced significant results: predicted scores were significantly higher for high than for low esteem subjects.

The major purpose of the present study was to examine the relationships between self-esteem measures at different levels of generality and academic performance. The first two-fold hypothesis was that global self-esteem (GSE) measures would be positively related to specific self-esteem (spSE) measures and that specific measures would be positively related to task-specific self-esteem (tsSE) measures; it was unclear whether or not GSE measures would be related to tsSE measures. Second, it was hypothesized that tsSE measures would be most closely related to scores on a midterm examination, that spSE measures would be less closely related, and that GSE measures would have the lowest degree of relationship, if any. Third, it was predicted that high esteem subjects would receive significantly higher grades than would moderate esteem subjects and that moderate esteem subjects would receive significantly higher grades than would low esteem subjects. Finally, it was hypothesized that spSE measures would have a stronger relationship than GSE measures to predicted performance.

### *Method*

Three different measures were used to assess global self-esteem. The first measure was the total score on the Tennessee Self-Concept Scale (TSCS) (Fitts, 1965) which is based on 90 items summed across three

internal frames of reference (identity, self-satisfaction, and behavior) each in relation to five aspects of the self (physical, moral-ethical, personal, familial, and social self). The second measure was Rosenberg's (1965) self-esteem test (RbSE) composed of 10 items that form a Guttman scale. The third measure was a single item (QSE) constructed for this study on which subjects rate their global self-esteem on a 10-point scale after comparing themselves to descriptions of high and low esteem persons provided as anchor points.

Two measures of specific self-esteem were also developed since neither the TSCS nor the RbSE has subscales relevant to academic performance. A rating of *intellectual esteem* (inE) was obtained by asking subjects, "Generally, how high is that part of your esteem which is based on your assessment and evaluation of your intellectual abilities?" This was judged to represent the complex set of skills most nearly pertinent to academic performance. A rating of *educational esteem* (edE) was obtained by asking subjects, "Generally, how high is your esteem in academic-educational situations (in your classes and other situations directly related to your education)?" This question was judged to represent the general area in which academic performance would fall. Ratings for both measures were made on a 10-point scale. Pilot work done with these two measures and on the QSE yielded two-month, test-retest reliabilities of .84, .81, and .77, respectively.

Two measures of task-specific self-esteem were obtained from predictions by subjects of their performance on a midterm examination. Before the test began, subjects were asked to estimate the numerical score they expected to receive on a standard academic scale (A = 90-100, B = 80-89, etc.). These estimates along with subjects' grade point averages (GPA) were collected before the tests were distributed. When the examination was over, subjects were again asked to predict their grades. These two estimates were designated task-esteem-before (tE-b) and task-esteem-after (tE-a).

Subjects were 78 male and 81 female students enrolled in a sophomore-level psychology course who completed all measures for the study. The TSCS and RbSE scales were administered in class during the third and fourth weeks of the quarter; the nine-item Self-Esteem Questionnaire was completed during the fifth week. The midterm examination, the first of two tests in the course, was given the sixth week during which time the two task-specific measures were obtained. The test consisted of 20 short-answer essay questions covering all lecture material and textbook reading assigned for the first half of the course. Raw scores were rescaled to conform to a standard academic scale and were used along with scores predicted before the test (the tE-b measure) as the dependent measures.

### Results

Independent  $t$  tests performed on all male-female comparisons revealed no significant differences for any of the nine measures except midterm grade: females did receive significantly higher scores than did males,  $t(157) = 3.28, p < .01$ . To check further for any sex differences, product-moment correlations were computed among all variables for each sex separately. Inspection of the two matrices revealed no major differences in either magnitude or pattern of the correlations; hence the data were combined for the remaining analyses.

The correlation matrix for all subjects was factor analyzed through using the principal components method with a varimax rotation, and two factors accounting for 52.7% and 47.3% of the variance were extracted. Table 1 presents intercorrelations and factor loadings for each of the nine measures. The two tsSE measures and the two academic measures had the highest primary loadings on the first factor; hence it was labeled *Academic Performance*. The high loadings of the three GSE measures clearly identified the second factor as *Global Self-Esteem*. The two spSE measures loaded about equally on these two factors.

Examination of Table 1 indicates that the patterns of relationships among the various esteem measures were as predicted. The mean correlation between the three GSE measures (.56) was similar to the correlation between the two spSE measures (.66) and to that between the two tsSE measures (.63), but slightly higher than the correlation between grade and GPA (.49). Correlations of measures within a subgroup with measures in other subgroups were all of about the same magnitude. Low but significant correlations were found between GSE and spSE measures ( $\bar{r} = .28$ ) as well as between spSE and tsSE measures ( $\bar{r} = .30$ ); however, the GSE measures were only marginally related to the tsSE measures ( $\bar{r} = .13$ ). Between each of the global, specific, and task-specific measures and the two academic measures the means of each set of resulting correlations were  $-.02$ ,  $.26$ , and  $.45$ , respectively. The only substantial difference in correlational indices involving self-esteem measures within a subgroup was the higher correlation of the tE-a measure with grade in comparison with the correlation between tE-b and grade (.54 vs. .41).

The distribution of scores for each esteem measure was divided into thirds to form high, moderate, and low esteem groups. Subjects whose scores fell on the dividing lines were randomly assigned to the appropriate group so as to assure equal size groups. Table 2 presents mean predicted grades and mean grades for each of these three esteem groups by esteem measure. A one-way analysis of variance performed

TABLE 1

*Intercorrelations and Factor Loadings for Esteem and Academic Measures*

Measure	Intercorrelation								Factor Loading	
	2	3	4	5	6	7	8	9	I	II
Global										
1. TSCS	.59	.50	.26	.30	.17	.04	.03	.00	.00	.79
2. RbSE		.58	.23	.26	.17	.10	.01	.00	.00	.81
3. QSE			.27	.37	.13	.16	-.09	-.04	-.01	.82
Specific										
4. inE				.66	.27	.27	.24	.30	.47	.55
5. edE					.32	.35	.23	.29	.52	.48
Task										
6. tE-b						.63	.41	.43	.74	.17
7. tE-a							.54	.42	.80	.09
Academic										
8. grade								.49	.76	-.10
9. GPA									.75	-.06

Note.—For  $n = 159$ ,  $p < .05$  for  $r \geq .16$ ;  $p < .01$  for  $r \geq .21$ .

on the midterm scores of the three groups revealed no significant effects for any of the GSE measures and only a marginally significant effect for the inE measure. However, significant effects were obtained for the edE, tE-b, and tE-a measures. A post-hoc comparison of means using Newman-Keuls tests (Winer, 1971) showed that the high esteem group did obtain significantly higher scores on the edE measure than did the moderate esteem group,  $q(3, 156) = 3.45$ ,  $p < .05$ , or than did the low esteem group,  $q(2, 156) = 2.82$ ,  $p < .05$ . For the tE-b measure, high esteem subjects did score significantly higher than did low esteem

TABLE 2

*Mean Predicted Grades and Mean Grades of High, Moderate, and Low Esteem Groups by Esteem Measure*

Esteem Groups by Esteem Measure								
Esteem Measure	Mean Predicted Grade Esteem Group <sup>a</sup>				Mean Grade Esteem Group <sup>a</sup>			
	High	Moderate	Low	F <sup>b</sup>	High	Moderate	Low	F <sup>b</sup>
TSCS	82.8	81.0	81.5	1.00	72.1	69.3	71.3	1.13
RbSE	83.2	81.5	80.8	1.51	72.5	70.8	71.9	1.00
QSE	83.3	81.5	81.6	1.00	70.3	70.9	72.8	1.04
inE	84.1	80.6	80.2	3.43**	73.3	70.0	69.8	2.77*
edE	83.7	82.0	79.6	4.64**	73.9	69.6	70.4	3.43**
tE-b					74.7	71.8	66.9	12.60***
tE-a					76.4	70.3	66.1	25.11***

<sup>a</sup>  $n$  per group = 53

<sup>b</sup>  $df = 2, 156$ .

\*  $p < .10$ .

\*\*  $p < .05$ .

\*\*\*  $p < .001$ .



subjects,  $q(3, 156) = 7.03, p < .01$ , and the moderate esteem subjects did score significantly higher than did the low esteem subjects,  $q(2, 156) = 4.39, p < .01$ . For the tE-a measure, the high esteem group scored significantly higher than both the low esteem sample,  $q(3, 156) = 10.02, p < .01$ , and the moderate esteem sample,  $q(2, 156) = 5.88, p < .01$ , whereas the moderate esteem group did register significantly higher scores than did the low esteem group,  $q(2, 156) = 4.14, p < .01$ .

A one-way analysis of variance performed on the predicted midterm scores of the high, moderate, and low esteem groups revealed significant effects for both spSE measures, but not for any of the GSE measures. For the inE measure, high esteem subjects did yield significantly higher predicted grades than did the moderate esteem subjects,  $q(2, 156) = 3.70, p < .05$ , or than did low esteem subjects,  $q(3, 156) = 4.10, p < .01$ . For the edE measure, higher grades were predicted for high esteem subjects than for the low esteem subjects,  $q(3, 156) = 4.29, p < .01$ .

### *Discussion*

The results of this study supported all four hypotheses. Significant correlations were found between global and specific measures and between specific and task-specific measures, but correlations between global and task measures were only of marginal significance. Task-specific measures did show a stronger relationship to academic performance than did specific measures which in turn had a stronger relationship to grades than did global measures. Although these findings are in agreement with Schrauger (1972), they point even more clearly to the predictive power of the task-specific measures. They also corroborate many of the relationships between variables found by Morrison et al. (1973) with one important exception—global self-esteem was unrelated both to predictions of performance and to actual performance.

To account for these findings it is necessary to examine the scales used to measure the different self-esteem constructs. At least four important criticisms have been made of global measures. First, Gergen (1971) has pointed out that global measures tend to overlook important situational influences. When a global measure such as the Self-Esteem Inventory (Coopersmith, 1967) represents the specific area pertinent to the behavior under study, then a relationship between global self-esteem and behavior may be obtained as in Morrison et al. (1973). However, when an instrument such as the Tennessee Self-Concept Scale does not represent a particular area or when it is not possible to determine how much subjects consider a particular area in



the completion of generally stated items such as those on the RbSE or QSE scales used in this study, significant relationships may not be obtained.

A second issue concerns the weighting of different sources of self-esteem (Rosenberg, 1965). One person may base his global self-esteem more on intellectual-academic achievements, whereas another may base it more on social or characterological factors. Most global measures, however, do not take these differences into account. The TSCS weights each of five areas equally, whereas differential weights on the RbSE and QSE scales are unknown, since subjects have the freedom to determine weights themselves.

A third difficulty stems from the fact that the global self-esteem construct is so all-encompassing that it is hard to know what is being measured (Wylie, 1974). Global measures may represent literally thousands of behaviors in a wide variety of situations, and yet these behaviors, as well as subjects' values, standards, reference groups, and evaluative criteria, are often unknown. These and other problems of measurement described by Wylie (1974) decrease the predictive power of many global measures.

In contrast, these criticisms do not apply so much to specific and task-specific measures as to global measures. Specific measures as contrasted with global measures can focus upon more restricted areas of functioning or categories of behavior. However, these specific measures are still rather broad because they represent, theoretically, the weighted sum of evaluations for hundreds of different behaviors and situations. Furthermore, the situational referents, behaviors, and weights for specific measures are unknown. Task-specific measures, on the other hand, can focus upon specific behaviors in a specific situation. Thus, when esteem level is related to specific behaviors, it is easy to see how more specific measures would yield higher correlations than would global measures.

A fourth criticism of self-esteem measures may also apply more to global measures than to specific measures. Most individuals are motivated to evaluate themselves positively (Rosenberg, 1965), a tendency which operates to bias their self-reports (Ziller et al., 1969). When subjects describe or evaluate themselves in a general setting or fashion rather than in a specific situation or than in relation to specific behaviors, they may be more likely to present themselves in a socially desirable, positive, or idealistic manner. However, when the evaluation process is restricted or when subjects are confronted with their own behavior in a specific situation, they may be less likely to furnish socially desirable self-reports than they would in the global circumstances. This positive distortion in ratings may be one factor that

accounts in part for the findings that global self-esteem is related to predictions of performance, but not to actual performance.

The findings of this study again question the utility of the global self-esteem construct and underscore the importance of several suggestions made by Wylie (1974). It is imperative that researchers first clearly define more specific kinds of self-esteem and then, making full use of the methodological expertise available, develop instruments to measure them. Global self-esteem measures should be developed or refined so that in representing major sources of self-esteem they may be weighted in terms of their importance to the subject. The construct validity of these instruments must then be established. Researchers using self-esteem measures would be well-advised to select the measure most nearly appropriate for their study: global measures may be preferable in some cases, but specific or task-specific measures may be of greater value in others.

## REFERENCES

- Coopersmith, S. *The antecedents of self-esteem*. San Francisco: W. H. Freeman, 1967.
- Fitts, W. H. *Tennessee Self-Concept Scale: Manual*. Nashville, Tenn.: Counselor Recordings and Tests, Department of Mental Health, 1965.
- Gergen, K. J. *The concept of self*. New York: Holt, Rinehart & Winston, 1971.
- Morrison, T. L., Thomas, M. D., and Weaver, S. J. Self-esteem and self-estimates of academic performance. *Journal of Consulting and Clinical Psychology*, 1973, 41, 412-415.
- Rosenberg, M. *Society and the adolescent self-image*. Princeton, New Jersey: Princeton University Press, 1965.
- Schrauger, J. S. Self-esteem and reactions to being observed by others. *Journal of Personality and Social Psychology*, 1972, 23, 192-200.
- Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1971.
- Wylie, R. C. *The self-concept: A review of methodological considerations and measuring instruments*. (Revised ed., Vol. 1). Lincoln: University of Nebraska Press, 1974.
- Ziller, R. C., Hagey, J., Smith, M. D., and Long, B. Self-esteem: A self-social construct. *Journal of Consulting and Clinical Psychology*, 1969, 33, 84-95.

## THE VALIDITY OF SOME ALTERNATIVE MEASURES OF ACHIEVEMENT MOTIVATION

FRANK B. W. HARPER

University of Western Ontario, Canada

Two tests commonly used in research on achievement motivation, the Thematic Apperception Test and the Test Anxiety Questionnaire, have been criticized for a number of shortcomings involving reliability, validity, and ease of scoring. The present study examines the retrospective validity of two alternate measures which appear to overcome many of the objections to the former tests. These alternate measures are the *n-Ach* scale of the Personality Research Form and the Debilitating Anxiety scale of the Achievement Anxiety Test. An analysis of variance comparing the relative academic achievement of two samples of college students was performed through using high and low scoring comparisons on the two measures. The results showed that academic achievement was significantly related to scores on the tests. The alternate measures are therefore recommended to researchers for further study of achievement motivation.

ACHIEVEMENT motivation theory as formulated by Atkinson (1964) has used the need achievement score (*n-Ach*) derived from the Thematic Apperception Test (TAT) as the principal operational measure of the motive for success *Ms*, and the Test Anxiety Questionnaire (TAQ) as the principal operational measure of the motive to avoid failure *M<sub>AF</sub>*.

Clarke's (1973) critique of the use of the TAT as a measure of need achievement and presumably therefore of *Ms*, and Harper's (1971, 1974) critique of the TAQ as a measure of test anxiety and, therefore, of *M<sub>AF</sub>*, raise questions about the reliability and validity of these measures. Alternative measures of need achievement which meet more stringent criteria of reliability than does the TAT are recommended by Clarke. In particular he has suggested the use of the need achievement

scale (*n-Ach*) of the Personality Research Form (PRF) (Jackson, 1967), as a more suitable instrument. The PRF was designed to minimize the effects of social desirability and acquiescence response sets in self-report personality inventories. Jackson reported reliabilities for the *n-Ach* scale of .72 to .86. A high scorer on this scale is described as an individual who aspires to accomplish difficult tasks, maintains high standards, and is willing to work towards distant goals; responds positively to competition; is willing to put forth effort to attain excellence. Correspondingly, a low scorer on this scale is described in opposite terms.

Harper's review of the comparative validity of the two principal measuring instruments for test anxiety, the TAQ and the Achievement Anxiety Test (AAT) (Alpert and Haber 1960), recommended that when anxiety about taking conventional college examinations or tests was the focus of the researcher's interests, the Debilitating scale (*Deb*) of the AAT be used (Harper 1974).

### Problem

It was the purpose of this study to ascertain the retrospective validity of one formulation of the achievement motivation hypothesis in a college population through the use of these two operational measures, the *n-Ach* scale of the PRF and the *Deb* scale. In reviewing Spielberger's work on the relationship of Manifest Anxiety to grade point average (GPA), (Spielberger 1962, Spielberger and Katzenmeyer, 1959), Atkinson (1964, p. 255) suggested that groups formed from individuals high or low in score distributions of *n-Ach* measures or anxiety measures, should differ significantly in their academic attainment as reflected in their overall grade point averages. Weiner (1972, pp. 195-209) presented an hypothesis derived from Atkinson's analysis. Weiner developed the formal conceptual terms of achievement theory as follows: when the motive for success ( $M_s$ ), which is operationalized as a score on a *n-Ach* measure, is greater than the motive to avoid failure ( $M_{AF}$ ), which is operationalized as a score on a measure of test anxiety, the individual should approach achievement-related activities such as college examinations in a positive way. Conversely, when  $M_s$  is less than  $M_{AF}$ , the individual should be more hesitant about approaching achievement-related activities and consequently should be less proficient in examinations.

The more sophisticated versions of achievement motivation theory add to this formulation additional variables involving subjective estimates of the probability of success and the incentive value of success which a retrospective study of this kind cannot calculate. The theory is



therefore tested in its simplest form as a measure of the possible effect on overall career grade point average of the relationship of  $M_S$  to  $M_{AF}$ .

### *Subjects and Procedure*

Two samples of college students were studied. One was a sample of 304 male graduates and the other a sample of 332 women graduates. Both samples were enrolled in a post-graduate teacher preparation program, after having completed the Bachelor's degree in an academic field. Each subject was given the PRF and the AAT to complete at the beginning of the teacher preparation year. The overall career grade point averages for the samples were calculated from their respective college transcripts.

### *Scoring and Analysis*

The subjects were assigned to one of four cells in a  $2 \times 2$  ANOVA design, according to their scores on the two scales, *n-Ach* and *Deb*. Subjects scoring higher than the median on both scales were assigned to the High-High cell. Subjects scoring lower than the median on both scales were assigned to the Low-Low cell. Subjects scoring above the median in *n-Ach* but below the median in *Deb* were assigned to the High-Low cell. Finally, subjects scoring below the median in *n-Ach* and above the median in *Deb* were assigned to the Low-High cell. The grade point averages for each subject were then entered into the ANOVA as the dependent variable, and the corresponding  $F$  values calculated in a conventional  $2 \times 2$  design.

### *Results*

The grade point averages for each cell of the  $2 \times 2$  distribution are shown in Table 1, for each sex separately.

The GPA values follow the sequence predicted by the research hypothesis. Subjects in the High Success (*n-Ach*) and Low Test Anxiety (*Deb*) cell have the highest GPA, whereas subjects in the Low Success (*n-Ach*) and High Test Anxiety (*Deb*) cell had the lowest GPA. Subjects in the other two cells had GPA's which were midway in value between the other two cells.

A  $2 \times 2$  ANOVA was performed on the table. The values of  $F$  are shown in Table 2. It can be seen in both samples that the values of  $F$  were significant for the main effects of *n-Ach* and *Deb*. For the male sample a significant interaction effect did exist, whereas for the women's sample a significant interaction was not present.



TABLE 1  
*Grade Point Averages in Groups Differing in Four Combinations of  
 n-Ach and Test Anxiety*

		Males $N = 304$ Success ( <i>n-Ach</i> measure)	
Anxiety ( <i>Deb</i> measure)	High	High	Low
	Low	65.38 68.16	64.59 65.00
		Women $N = 332$ Success ( <i>n-Ach</i> measure)	
Anxiety ( <i>Deb</i> measure)	High	High	Low
	Low	68.48 69.23	66.34 67.89

### Discussion

The study was intended to ascertain the retrospective validity of two alternative measures of achievement motivation. The results show that in both samples, the alternate measures had significant relationships with career grade point average. Both the *n-Ach* scale of the PRF and the *Deb* scale of the AAT appear therefore to be worthy of further study as operational measures of the constructs  $M_s$  and  $M_{AF}$ , respectively. The ease of administration and scoring of these two measures compared to the difficulties in administering and scoring the TAT and the TAQ should recommend the use of them to researchers.

TABLE 2  
*Analysis of Variance of Grade Point Averages in Groups  
 Differentiated in Achievement Motivation*

Source	df	Mean Square	F	P
<i>Males</i>				
Test Anxiety ( $M_{AF}$ )	1	192.64	9.07	<.01
<i>N-Ach</i> ( $M_s$ )	1	296.05	13.95	<.01
Interaction	1	105.68	5.02	<.05
Within	300	21.22		
Total	303			
<i>Women</i>				
Test Anxiety ( $M_{AF}$ )	1	109.88	4.85	<.05
<i>N-Ach</i> ( $M_s$ )	1	251.56	11.10	<.01
Interaction	1	13.52	0.60	ns
Within	328	22.67		
Total	331			

## REFERENCES

- Alpert, R. A. and Haber, R. N. Anxiety in academic achievement situations. *Journal of Abnormal and Social Psychology*, 1960, 61, 207-215.
- Atkinson, J. W. *An introduction to motivation*. Princeton, N.J.: D. Van Nostrand, 1964.
- Clarke, D. E. Measures of achievement and affiliation motivation. *Review of Educational Research*, 1973, 43, 41-51.
- Harper, F. B. W. Specific anxiety theory and the Mandler-Sarason Test Anxiety Questionnaire. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1971, 31, 1011-1014.
- Harper, F. B. W. The comparative validity of the Mandler-Sarason Test Anxiety Questionnaire and the Achievement Anxiety Test. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1974, 34, 961-966.
- Jackson, D. N. *Personality research form manual*. New York: Research Psychologists Press Inc., 1967.
- Spielberger, C. D. The effects of academic anxiety on the academic achievement of college students. *Mental Hygiene*, 1962, 46, 420-426.
- Spielberger, C. D. and Katzenmeyer, W. C. Manifest anxiety, intelligence and college grades. *Journal of Consulting Psychology*, 1959, 23, 278.
- Weiner, B. W. *Theories of motivation*. Chicago: Markham Publishing Company, 1972.



## RELATIONSHIPS AMONG FOUR MEASURES OF ACHIEVEMENT MOTIVATION

THOMAS R. WOTRUBA  
Department of Marketing  
San Diego State University

KARL F. PRICE  
Department of Management  
Temple University

Short, objective tests of achievement motivation have considerable potential appeal to investigators for reasons of economy and ease of analysis. Two new paper-and-pencil tests of achievement motivation developed by Hermans and Mehrabian were examined to determine whether they might be comparable to two older measures, McClelland's TAT *n-Ach* and the achievement scale of the Edwards Personal Preference Schedule. The four measures were administered to 65 undergraduate business administration students at San Diego State University. Although the results reflected a modest (.30) correlation between the Hermans and the TAT *n-Ach*, no other significant correlations among pairs of the four achievement measures were found. The results lend support to past findings; namely, that the various achievement measures would appear to be measuring dissimilar constructs.

THE purpose of the study was to investigate whether significant relationships exist among four measures of need achievement. Two of these, McClelland's projective measure and the achievement (*ach*) scale of the Edwards Personal Preference Schedule (EPPS), have been subject to considerable previous study. The other two measures have been obtained from relatively new paper-and-pencil tests of achievement motivation. Developed by Mehrabian (1968, 1969), the first one includes "verbal items which are designed to discriminate high versus low achievers," (1968, p. 494). Separate male and female scales were devised, each with 26 items to be rated on a 9-point measure of agreement. Developed by Hermans (1970), the second one consists of

29 multiple choice items representing various aspects of achievement motivation found in the literature.

Short, objective tests of achievement motivation have considerable potential appeal to investigators because of relative economy and ease of analysis. However, numerous studies of the relationships between various measures of achievement motivation have produced little in the way of positive relationships. For example, the achievement score (*n-Ach*) of McClelland's Thematic Apperception Test (TAT) showed no significant relationship with three objective tests: Survey of Study Habits and Attitudes (SSHA), Opinion, Attitude, and Interest Survey (OAIS), and a 99 item How To Study test (HTS) (Krumboltz and Farquhar, 1957). In another study, the same TAT measure of *n-Ach* showed no significant correlation with self-reports or with self-peer ranking measures (Holmes and Taylor, 1968). A number of studies have failed to produce a significant correlation between the TAT *n-Ach* score and the achievement scale measure on the EPPS (Himelstein, Eschenbach, and Carp, 1958; Marlowe, 1959; and Melikian, 1958).

Thus, in addition to replicating previous studies concerning relationships between McClelland's *n-Ach* scores and the EPPS achievement scales, the present study adds information from two recently devised instruments to the pool of correlation data on achievement tests.

### Procedure

The subjects were 65 undergraduate business administration students at San Diego State University. The McClelland TAT was administered first; the EPPS one week later; and the two scales by Mehrabian and Hermans, about a week later, two days apart. Preceding all the tests, a demographic questionnaire was completed by each participant.

The TAT's were scored by both authors (Atkinson, 1958), with an interrater reliability of .88. The other three tests were scored according to procedures devised by each test's author (Edwards, 1959; Hermans, 1970; and Mehrabian, 1968, 1969).

### Results

The results, as shown in Table 1, reflect a modest but significant correlation between the TAT *n-Ach* and Hermans achievement measures. None of the other five pairs of achievement measures produced a degree of relationship significant at the .05 level, although the correlation between the Hermans and Mehrabian measures could occur by chance at the .10 level. Nevertheless, both the Hermans and the Mehra-



TABLE 1  
Correlations between Study Variables

EPPS variables	TAT measure of <i>n-Ach</i>	Hermans	Mehrabian
<i>ach</i>	-.169	.219	.215
<i>def</i>	-.018	-.149	-.010
<i>ord</i>	.076	.090	-.364**
<i>exh</i>	-.123	.001	.026
<i>aut</i>	.000	-.192	-.094
<i>aff</i>	.059	-.262*	-.029
<i>int</i>	.020	.082	.108
<i>suc</i>	.072	-.238	-.391**
<i>dom</i>	-.127	.291*	.343**
<i>aba</i>	.107	-.025	-.067
<i>nur</i>	.189	-.090	.148
<i>chg</i>	-.145	-.079	.142
<i>end</i>	.205	.318**	.076
<i>het</i>	.018	-.218	-.094
<i>agg</i>	-.224	.123	-.059
<i>cons</i>	-.034	.190	.296*
Mehrabian	.134	.234	
Hermans	.299*		

\*  $p \leq .05$ .\*\*  $p \leq .01$ .

bian measures correlated significantly with at least three of the EPPS variables other than achievement. Both shared a significant common variance with dominance (*dom*), and either one or the other showed a significant negative correlation with order (*ord*), succorance (*suc*), affiliation (*aff*), and a significant positive correlation with endurance (*end*).

When achievement measures are related to the subjects' demographic and classification data, only four out of 28 pairs of relationships emerged with low but significant ( $p \leq .05$ ) correlation coefficients. (See Table 2.)

TABLE 2  
Relationship to Demographic Data

Achievement Measure	Demographic or Classification Measure	Correlation Coefficient
EPPS <i>ach</i>	grade point average	.276
Hermans	grade point average	.278
TAT <i>n-Ach</i>	age	-.245
TAT <i>n-Ach</i>	socio-economic group	.274

### Conclusion

In general, although a few small correlations were found in this investigation, only one such relationship occurred between the achievement tests themselves. Thus, the evidence presented continues to support the general thrust of past findings, namely that the various achievement motivation measures are in fact measuring dissimilar constructs. Furthermore, the new paper-and-pencil tests, the Hermans and the Mehrabian, are apparently not valid substitutes for each other or to any large extent replacements for other achievement measurement techniques.

### REFERENCES

- Atkinson, J. W. (Ed.). *Motives in fantasy, action, and society*. Princeton, N.J.: Van Nostrand, 1958.
- Edwards, A. L. *Manual-Edwards Personal Preference Schedule*. (Rev. ed.) New York: Psychological Corporation, 1959.
- Hermans, H. J. M. A questionnaire measure of achievement motivation. *Journal of Applied Psychology*, 1970, 54, 353-363.
- Himelstein, P., Eschenbach, A. E., and Carp, A. Interrelationships among three measures of need achievement. *Journal of Consulting Psychology*, 1958, 22, 451-452.
- Holmes, D. S. and Tyler, J. D. Direct versus projective measures of achievement motivation. *Journal of Consulting and Clinical Psychology*, 1968, 32, 712-717.
- Krumboltz, J. D. and Farquhar, W. W. Reliability and validity of the n-Achievement test. *Journal of Consulting Psychology*, 1957, 21, 226-228.
- Marlowe, D. Relationships among direct and indirect measures of achievement motivation and overt behavior. *Journal of Consulting Psychology*, 1959, 23, 329-332.
- Mehrabian, A. Male and female scales of the tendency to achieve. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1968, 28, 493-502.
- Mehrabian, A. Measures of achieving tendency. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1969, 29, 445-451.
- Melikian, L. H. The relationship between Edwards' and McClelland's measures of achievement motivation. *Journal of Consulting Psychology*, 1958, 22, 296-298.

## PREDICTION OF PERSISTENCE AND PERFORMANCE WITH THE HERMANS PRESTATIC MOTIVATION TEST<sup>1</sup>

J. OGDEN HAMILTON<sup>2</sup>

Indiana University

Hermans' Prestatic Motivation Test, a questionnaire measure of achievement motivation, is easier to administer and to score than are its projective counterparts, the Thematic Apperception Test and the French Test of Insight; and it need not be administered under controlled conditions. In two independent studies of its predictive validity, Hermans' measure was found to be positively related to persistence and to performance in academic examinations, both when the measure was used alone and when it was combined with the Mandler-Sarason Test Anxiety Questionnaire as a measure of resultant motivation. Moreover, although the French Test of Insight was found to be related to persistence as in earlier research, it was not related to Hermans' measure. It is concluded that Hermans' questionnaire taps a psychological characteristic that is manifest in achievement directed behavior, but that this characteristic is something other than the achievement motive of McClelland and his colleagues.

THE purpose of this investigation was to examine the degree of relationship between achievement motive measured by Hermans' Prestatic Motivation Test (PMT) and measures of persistence and performance in academic tasks. The measure, which consists of 29 Guttman-scaled items, amounts to a self-report measure of attitudes and behaviors previously shown to be related to the achievement motive. It

<sup>1</sup> This research was partially funded by a grant from the Division of Research, Graduate School of Business, Indiana University. Parts of this research were presented at the annual meetings of the American Psychological Association, New Orleans, 1974.

<sup>2</sup> The author acknowledges the editorial contributions of James A. Wall, Jr. Requests for reprints should be sent to J. Ogden Hamilton, Graduate School of Business, Indiana University, Bloomington, Indiana 47401.

was described in detail by Hermans (1970). Its principal strength is logistical: As a questionnaire, it does not require the controlled conditions or the scoring expertise needed to use the best known projective measures of the achievement motive—the Thematic Apperception Test (TAT) (McClelland, Atkinson, Clark, and Lowell, 1953) and the French Test of Insight (FTI) (French, 1958). Because this flexibility is valuable, even indispensable in field research, the PMT seemed to merit validation beyond that initially reported by Hermans. This study was undertaken to ascertain whether the PMT predicts the same sorts of behavior as do the projective measures. If it does, then one can be more confident than previously in designing and interpreting future research with the measure on the basis of the wealth of knowledge about the achievement motive accumulated by McClelland and his followers.

To maintain consistency with the motivation literature, the PMT was investigated both alone and in combination with the Mandler-Sarason Test Anxiety Questionnaire (TAQ) (Mandler and Sarason, 1952), a measure of fear of failure. The TAQ often has been used with each of the common projective measures to tap a joint construct, resultant motivation. In theory, the more the achievement motive of an individual exceeds his fear of failure, the higher his resultant motivation and the more likely he is to engage in achievement seeking behavior. Atkinson and Litwin (1960) have demonstrated that a measure of resultant motivation can be a more accurate predictor of behavior than is achievement motive alone. In the present studies, resultant motivation was measured by subtracting the rank on the TAQ from the rank on the measure of achievement motivation, a high rank always indicating a high score on a measure.

### *Study I: Persistence*

As an accepted behavioral manifestation of the achievement motive, persistence has been related to both common projective measures of the motive: the TAT by Feather (1961) and the FTI by Atkinson and Litwin (1960). Since the present study was intended simply to determine whether the PMT could be substituted for the projective measures used in earlier studies, the two-fold hypothesis offered was the same one as in these earlier studies: namely, that persistence in a difficult task would be positively related to the achievement motive measured by the PMT as well as to resultant motivation measured by the PMT and the TAQ.

### Method

The 41 subjects were male undergraduates in a behavioral science course. All the data were collected in connection with the pedagogical requirements of the course. To permit direct comparison of the PMT with one of the standard projective measures, the FTI was administered in addition to the PMT and the TAQ. The FTI was scored by two experienced specialists who achieved satisfactory reliability ( $r_s = .88$ ).

The PMT consisted of the 29 items used by Hermans (1970). The measure was scored as described by Hermans: i.e., those above the median on each item received a score of one for that item, and those below, a zero. Contrary to Hermans' procedure, those scoring at the median were given a fractional score derived from the proportional allocation of the median rather than an arbitrary zero or one:

$$\text{Fractional score} = 1 - \frac{n/2 - \text{no. below median}}{\text{no. tied at median}} \quad (1)$$

This procedure is superior to Hermans' scoring, for it is more sensitive to the assumption of a theoretically continuous distribution of scores on each item. In effect, the modification calculates the likelihood that any given observation tied at the median in fact lies above the true median.

Persistence was measured by the length of time spent in the course final examination, with no time limit. This is the measure that was used by Atkinson and Litwin (1960), the assumption being that the higher the achievement motive of a student, the longer he spends attempting to do well on the examination. When a student finished the final examination and left the room, the experimenter noted the time on a card and asked the student to enter his social security number on the card.

Because a student's verbal ability reasonably might be related to the length of time he spends on an examination, scores on the Verbal section of the Scholastic Aptitude Test were held constant statistically during the data analysis.

### Results

As predicted, resultant motivation based on the PMT was significantly related to persistence ( $r = .28, df = 38, p < .05$ ). The relationship between persistence and the PMT alone was not quite so strong



as that between persistence and the measure of resultant motivation ( $r = .21$ ,  $df = 38$ ,  $p < .10$ ). The relationships between persistence and both the FTI and the corresponding measure of resultant motivation also were positive, but neither achieved the chosen level of significance. The PMT and the FTI were virtually uncorrelated ( $r = .07$ ).

### *Reanalysis and Results*

The lack of a significant relationship between the FTI and the time spent in the examination prompted investigation of the validity of the scores on the FTI. The original Atkinson and Litwin study used a less conservative analysis than the one just described, in that it involved only subjects with extreme motive strengths. Specifically, subjects scoring both above the median in achievement motive and below the median in fear of failure (i.e., high resultant motivation) were compared to those scoring both below the median in achievement motive and above the median in fear of failure (i.e., low resultant motivation). Subjects scoring above the median on both measures or below on both were eliminated. Reanalysis of the data of the present study using this procedure resulted in a precise replication of the original study and put the FTI onto its home ground, so to speak.

Using this procedure, it was found that both the PMT and the FTI showed the predicted relationship to persistence. For the PMT, the 10 subjects classified high in resultant motivation spent an average of 99 minutes in the examination, compared to 76 minutes spent by the 10 classified low in resultant motivation ( $t = 3.07$ ,  $df = 18$ ,  $p < .01$ ). For the FTI, the 11 subjects classified high in resultant motivation spent 106 minutes, compared to 78 for the 11 classified low in resultant motivation ( $t = 2.30$ ,  $df = 20$ ,  $p < .05$ ).

### *Study II: Performance*

Performance as well as persistence is an accepted behavioral manifestation of the achievement motive. Among others, Atkinson and Litwin (1960) have demonstrated that performance in an examination is positively related to the FTI, and Hermans (1970) has reported similar results with the PMT in his use of Dutch subjects. The two-fold hypothesis of the present study was the same one as in these earlier studies: namely, that performance in an achievement task would be positively related to the achievement motive measured by the PMT as well as to resultant motivation measured by the PMT and the TAQ.

### *Method*

The 24 subjects were students in an undergraduate behavioral science course. The specific class was chosen because of the instructor's grading policy. Prior to taking an examination students were given a list of possible questions from which the questions used on the examination were selected. Thus, the course grade for a student depended on the effort he was willing to devote to preparing answers to all the possible questions, and intelligence and verbal ability were less important than often would be the case. The course final grade was the measure of academic achievement.

The course instructor administered the PMT and the TAQ in connection with the pedagogical requirements of the course. He was not aware that the score from these tests would be used for research; moreover, he never saw the test scores. No projective measure was administered. To control for verbal ability, the Verbal scores on the Scholastic Aptitude Test were held constant statistically during the analysis.

### *Results*

As predicted, the score on the PMT was significantly related to performance ( $r = .54$ ,  $df = 21$ ,  $p < .005$ ). The relationship between performance and resultant motivation was not quite so strong as that between the PMT alone and performance ( $r = .29$ ,  $df = 21$ ,  $p < .10$ ).

### *Discussion*

These two studies suggest that the PMT indeed measures a psychological characteristic that manifests itself in achievement-seeking behavior. At this point, however, it is not clear whether the corresponding measure of resultant motivation is a more accurate predictor of behavior than is the PMT alone, for the data in these studies did not permit definitive comparison of the two measures. A clear determination of this issue requires further research.

On the basis of the reanalysis of the persistence data, in which the procedure of the original Atkinson and Litwin study of the achievement motive and persistence was used, it seems reasonable to believe that the FTI was validly administered and scored, and that it was related to persistence in the same way as it was in the Atkinson and Litwin study. Since the PMT did show nearly the same relationship with the same kind of behavior of the same sample as did the FTI, it is reasonable to believe that it, too, would be a valid measure of a

predisposition to achieve. Therefore the very low correlation between the two measures can be interpreted with some confidence as indicating that although each measure taps a psychological characteristic that is manifest in achievement-seeking behavior, the two measures do not represent the same characteristic.

In sum, it is concluded that the PMT promises to be a valuable tool for studying achievement-related behavior, especially when use of the standard projective measures is impractical. However, it must be realized that since the PMT seems not to measure the achievement motive of McClelland and his colleagues, one should not make indiscriminate use of those aspects of the achievement motive literature that have not been explicitly studied with the PMT. The present studies have at once shown the promise of the measure and also the critical importance of further step-by-step establishment of its construct validity.

### REFERENCES

- Atkinson, J. W. and Litwin, G. H. Achievement motive and test anxiety conceived as motive to approach success and motive to avoid failure. *Journal of Abnormal and Social Psychology*, 1960, 60, 52-63.
- Feather, N. T. The relationship of persistence at a task to expectation of success and achievement-related motives. *Journal of Abnormal and Social Psychology*, 1961, 63, 552-561.
- French, E. G. Development of a measure of complex motivation. In J. W. Atkinson (Ed) *Motives in fantasy, action, and society*. Princeton: Van Nostrand, 1958.
- Hermans, H. J. M. A questionnaire measure of achievement motivation. *Journal of Applied Psychology*, 1970, 54, 353-363.
- Mandler, G. and Sarason, S. B. A study of anxiety and learning. *Journal of Abnormal and Social Psychology*, 1952, 47, 166-173.
- McClelland, D. C., Atkinson, J. W., Clark, R. W., and Lowell, E. L. *The motive to achieve*. New York: Appleton-Century-Crofts, 1953.

## A PRELIMINARY VALIDATION OF AN INSTRUMENT TO MEASURE THE DEGREE OF COUNSELOR RESTRICTIVE- NONRESTRICTIVE COGNITIVE ORIENTATION

THOMAS A. SEAY AND F. TERRILL RILEY

Kutztown State College

The present study was designed to provide a source of empirical validity for an instrument which measures the degree of cognitive functioning of a counselor along a restrictive-nonrestrictive dimension. The restrictive-nonrestrictive dimension refers to a holistic orientation as a mode for experiencing life by the receptivity toward the processing of and responding to sources of internal and external stimuli. To establish validity for the instrument, one hundred eleven counselor trainees in different phases of a training program designed to produce open and humanistic counselors were compared on the Counselor *R* Scale. The Rokeach Dogmatism Scale was included for analysis, since it was thought to be a component of the restrictive-nonrestrictive dimension. In the use of a  $2 \times 4$  analysis of variance design for unequal *n*'s, the study provided data supporting the hypothesis that counselors in different phases of their training would differ in their scores on the Counselor *R* Scale. As a trainee progresses through a humanistically oriented training program, he or she can be expected to move from the restrictive to the nonrestrictive ends of the measured dimension. Conclusions, implications, and future research potential were described.

In a comprehensive review of the theoretical and research literature on the characteristics of effective counselors, Shertzer and Stone (1968) concluded that "at the present time, the counseling profession is unable to demonstrate consistently that a single trait or pattern of traits distinguishes an individual who is or will be a 'good' counselor" (pp. 170-171). In a similar review, Brammer and Shostrom (1968) reached the same conclusions. Thus, although numerous characteristics have been identified, little consistency exists in the empirical

conclusions to justify their use in the selection or training of counselors. However, a re-examination of the research literature indicates that such pronouncements may be premature.

If the empirical findings are reordered using a different perspective, a single domain or dimension of traits emerges that could be characterized as a holistic orientation toward cognitively arranging and rearranging one's processed internal and environmental stimuli. Such a dimension amplifies what the cognitive and affective receptivity of a counselor to his encounter with life is and indicates how others perceive and respond to that encounter. The counselor's receptivity is thought, at this point, to originate from a developmental, physiological cognitive processing system.

If a pattern of characteristics emerges from the literature which allows conceptualization by means of a single dimension, then it becomes necessary to devise an instrument to assess the parsimony of behaviors coterminous with the conceptualization. The Counselor *R* Scale is an instrument which is intended to measure a restrictive-nonrestrictive dimension of counselor functioning. The dimension refers to the counselor's cognitive set or orientation for perceiving and processing organismic and environmental stimuli. The cognitive set leads to behaviors which indicate the degree to which the counselor is open and receptive toward people, things, and events that enter the counselor's frame of reference.

The cognitively nonrestrictive counselor as compared with one who is a restrictive counselor will tend to be more open and flexible in processing incoming stimulus events and, thereby, not only will remain open in his receptivity but, in addition, will actually expand his cognitive substructures. This cognitive activity, in turn, influences his behavior toward a nonrestrictive or appropriately restrictive mode of behaving. The cognitively restrictive counselor will tend to be just the reverse. In both instances, the behavioral mode will be reflected in the counselor's counseling orientation, approach, relationship, and use of selected counseling skills.

The present study sought to validate the Counselor *R* Scale as a measure of the dimension described. Because of the theoretical nature of the dimension, it should reflect changes in counselors-in-training as a result of entering a graduate program which emphasizes self and professional development toward a humanistic, open approach to counseling. Consequently, it was hypothesized that the *R* scale would reflect entry and progression through such a program. Of secondary interest was a potential male-female difference on the *R* scale.

An additional emphasis in the validation of the *R* scale was its correlation with the Rokeach Dogmatism Scale. Previous theorizing



and research by Rokeach (1960) indicated that dogmatism should be a component in the present dimension. Based on the previously stated expectation three Dogmatism items were found to be consistent with the present dimension and were included in the final construction of the *R* scale. Thus, because of expectations from theory and research and because of the inclusion of Dogmatism items, there should be a small positive correlation between the two scales.

### *Method*

#### *Subjects*

The subjects for the present experiment were 111 graduate students enrolled in various stages of a counselor education program at a small eastern college. The stages of program participation and, thus, the identification of experimental groups corresponded to four divisions derived from the number of graduate credit hours completed in the program. The four subgroups were designated as follows: (a) *Admitted Students*, accepted to the program but not enrolled in courses; (b) *Beginning Students*, with three to 12 credit hours earned in "core" courses; (c) *Mid-way Students*, with 15 to 24 credit hours completed; and (d) *Finishing Students*, with 27 to 39 credit hours accumulated. The subjects were heterogeneous in composition as might be expected in a counselor preparation program. Heterogeneity was maintained within each of the four groups and across those subjects who also had completed the Dogmatism instrument ( $n = 71$ ).

#### *Instrumentation*

*Counselor R Scale.* Since a theoretical description was presented previously, only the essential characteristics of the scale are discussed. Construction of the *R* scale followed the recommendations of Nunnally (1967) for developing an index of internal test validity through the unidimensionality of the construct being measured. Coefficient alpha for the *R* scale produced an internal consistency index of .84 ( $n = 107$ ), more than sufficient to suggest the existence of a relatively homogeneous measure of the construct or complex of constructs. Respective test-retest estimates of reliability of .84 ( $n = 64$ ) and .76 ( $n = 60$ ) were obtained over a two-week period and, subsequently, over a one to four month period. The *R* scale uses a Likert-type scale or continuum in which high scores are indicative of restrictiveness and low scores reflect nonrestrictiveness. The mean and standard deviation for the norm group were found to be 124.62 and 23.52, respectively.

The skewness and kurtosis values of .15 and  $-.53$ , respectively, suggested the presence of a relatively normal distribution of scores.

*Rokeach Dogmatism Scale.* The Rokeach (1960) scale is intended to measure the belief system of an individual. Because of its higher reported reliability ( $r = .91$ ), Version D (66 items) rather than another form was used in the present study.

### *Description of the Program*

The present validity study rests on the assumption that the counselor preparation program would enable students to move toward an open, humanistic life style orientation in counseling. To achieve that goal, the program was composed of two components: (a) self-development through self-awareness; and (b) professional skills development through competency-based preparation.

The self-development component was sought through formal and informal self-analysis and through systematic departmental faculty analysis. The primary purpose of the assessment procedures was to identify characteristics associated with effective and ineffective counseling which would enable the students by self-direction and faculty assistance to remediate weaknesses and to build on strengths.

Program components for professional skills development included both didactic and experiential learning with emphasis on development of a knowledge base, personal relationships, verbal and nonverbal skills, and total counseling strategies. For purposes of both self-development and professional development, much emphasis was placed on tape analysis during all phases of the program of study. The ultimate goal was the fully functioning counselor who would be cognitively aware of, receptive toward, and adaptive to a state of experiencing and who would have the ability to develop that state of experiencing as a tool for effective counseling. The departmental members, all of whom have been recipients of doctorates from major universities, professed a humanistic orientation, although different schools of thought have been represented.

### *Research Design*

The data for the preliminary validity study were analyzed through using a  $2 \times 4$  analysis of variance design for unequal  $n$ 's (Winer, 1962). Selected relationships between the means were analyzed by the Scheffe method of using orthogonal contrast coefficients for a priori data analysis (Winer, 1962, pp. 88-89). The correlation coefficient between the Dogmatism and the  $R$  scale was the Pearson  $r$  (Nunnally, 1967).

TABLE 1  
Mean Raw Scores for Female and Male Students in Four  
Positions within a Counselor Education Program

	Subgroup 1 Admitted Students	Subgroup 2 Beginning Students	Subgroup 3 Mid-Way Students	Subgroup 4 Finishing Students	
Female	136.15	119.75	116.25	109.37	121.54
Male	145.16	122.85	141.37	109.57	131.31
	141.38	121.08	128.81	109.46	

### Results

The analysis of variance summarized in Table 2 reveals that there were significant differences among the means of the eight subsamples formed by the inclusion of counselor sex with each of four positions of participation in a counselor education program,  $F(7, 103) = 5.99, p < .001$ . On the other hand, the differences between means associated with sex of the subjects (Factor A) reached significance at slightly less than the .05 level, but not at the .01 level,  $F(1, 103) = 4.09, p < .05$ . Differences between means as a function of program position (Factor B) attained significance considerably beyond the .001 level,  $F(3, 103) = 10.19, p < .001$ . However, the interaction variance between Factors A and B failed to reach significance,  $F(3, 103) = 1.47, p > .05$ .

Table 3 reports the a priori analysis of selected differences between the mean raw scores for the four program positions. The a priori assignment of orthogonal contrast coefficients reveals that the comparisons and results were as follows:

1. The difference between the mean of the subgroup of *Finishing Students* and the composite mean of the remaining three subgroups of students in different positions was statistically significant,  $F(1, 103) = 13.62, p < .001$ .

TABLE 2  
Analysis of Variance for Counselor R Scale Raw Scores

Source of Variance	SS	df	MS	F	P
Between all Subsamples	15,996.21	7	2,285.17	5.99**	.000007
A (Sex)	1,564.10	1	1,564.10	4.09*	.04
B (Position)	11,658.69	3	3,886.23	10.19**	.000006
A × B (Interaction)	1,690.12	3	563.37	1.47	.22
Within all Subsamples	39,254.97	103	381.11		

\* Significant at  $p < .05$ .  $F_{.05}(1, 103) = 3.91$

\*\* Significant at  $p < .01$ .  $F_{.01}(7, 103) = 2.79$ ;  $F_{.01}(3, 103) = 3.95$ ;  $F_{.01}(1, 103) = 6.85$ .

TABLE 3  
*Orthogonal Contrast Among Means in Relation to Four Positions  
 (Factor B) in the Counselor Education Program*

Designated Contrast	MS	F	P
$\psi_1 (\mu_4 - \mu_1 + \mu_2 + \mu_3 = 0)$ 3	5190.92	13.62**	.0003
$\psi_2 (\mu_3 - \mu_2 + \mu_4 = 0)$ 2	5098.82	13.38**	.0004
$\psi_3 (\mu_2 - \mu_3 = 0)$	719.88	1.89	.1700
Within variance	381.11		

\*\*  $p < .001$ ;  $F_{.05} (1, 108) = 6.85$ .

2. The difference between the mean of the subgroup of *Admitted Students* and the combined mean of the two subgroups of *Beginning Students* and *Mid-Way Students* was statistically significant,  $F(1, 103) = 13.38, p < .001$ .
3. The simple difference between the mean of the two subgroups of *Beginning Students* and *Mid-Way Students* was not statistically significant,  $F(1, 103) = 1.89, p > .05$ .

The Rokeach Dogmatism Scale correlated .31 with the *Counselor R Scale*. The percentage of variance common to both instruments was 9.6% ( $r^2 = .096$ ). Thus, 90.4% of the variance was unaccounted for or unexplained.

### Discussion and Conclusions

The analysis of the data confirmed the hypothesis that the instrument would reflect behaviors typically expected of open, humanistic counselors as differentiated by levels of their training program. Subjects who had not yet begun their program of study tended to be more restrictive and thus, less open and flexible in their mode of behaving as measured by the *R* scale than were members of the remaining three subgroups. The *Beginning* and *Mid-Way* subgroups were found to be similar in the degree of restrictive-nonrestrictive behavior. This last finding could be explained somewhat from the comparatively high score of the males in the *Mid-Way* subgroup. The less restrictive subgroup was the one composed of those subjects who were completing their training. In addition, if the two middle subgroups are combined, as indicated by the nonsignificant *F* test, an approximate linear trend from restrictive to nonrestrictive behavior styles of responding can be identified where the subgroup of *Admitted Students* was the most restrictive and the subgroup of *Finishing Students* was the

least restrictive. Thus, the *Finishing Students* who would be expected to demonstrate the more cognitively flexible processing system, as well as the more open, receptive attitude toward whatever they had been processing, did score on the average toward the nonrestrictive end of the dimension.

Important in relation to the theoretical meaning assigned to the *R* scale was the lack of a highly significant male-female difference on the restrictive-nonrestrictive dimension. Theoretically, it might be possible to infer that the male and female counselors studied tended to structure their perceptions in cognitively similar ways. In terms of utility, the *R* scale appears to be almost equally viable for male and female counselors.

Only 9.6% common variance was shared by the Dogmatism Scale and the *R* scale. Thus, although each scale appeared to contain a small component of the other, each was a unique measure of a different construct. The finding substantiated expectations.

Since the findings of the present study support the *R* scale as a measure of counselor behavioral restriction and by inference as an indicator of cognitive restriction, it now becomes possible to link the restrictiveness-nonrestrictiveness dimension to other behaviors in counseling and in other social relationships. In addition, the dimension should be explored in terms of selection, training, and selected counselor and client characteristics which determine effectiveness. It is anticipated that the nonrestrictive counselor would be the more effective counselor and that during the counseling process the cognitively nonrestrictive counselor is adaptive enough to be both nonrestrictive and appropriately restrictive. The parameters and implications of this concept must be examined. Ultimately, the *R* scale should be connected to the way an individual cognitively structures his internal world. Finally, the viability of the instrument for different populations such as teachers and ministers should be investigated.

In conclusion, the Counselor *R* Scale can be said to have face, content (unidimensionality), and empirical validity. It now becomes possible to attempt to determine the limits of its usefulness as a measure of counselor characteristics and, at the same time, expand the empirical validity as an instrument which measures cognitive functioning.

## REFERENCES

- Brammer, L. M. and Shostrom, E. L. *Therapeutic psychology*. Englewood Cliffs, New Jersey: Prentice-Hall, 1968.
- Nunnally, J. C. *Psychometric theory*. New York: McGraw-Hill Book Co., 1967.



- Rokeach, M. *The open and closed mind*. New York: Basic Books, 1960.
- Shertzer, B. and Stone, S. C. *Fundamentals of counseling*. Boston: Houghton Mifflin Co., 1968.
- Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill Book Co., 1962.

## HIGH SCHOOL TYPE, SEX, AND SOCIO-ECONOMIC FACTORS AS PREDICTORS OF THE ACADEMIC ACHIEVEMENT OF UNIVERSITY STUDENTS

JOHN F. McDONALD

University of Illinois at Chicago Circle

MICHAEL S. McPHERSON

Williams College

Grade point average was predicted for a sample of 152 students in Principles of Economics classes at the University of Illinois at Chicago Circle. It was shown that knowledge of high school type, sex, number of credit hours taken, and perhaps dollar value of scholarships and number of hours of outside work could significantly increase the ability to predict grades beyond that accomplished through using rank in high school class and American College Testing Program (ACT) Composite Score.

IN recent years many studies have been undertaken with the purpose of finding variables which increase the predictability of college grades beyond that accomplished by using a measure of high school performance and a score on a standardized scholastic aptitude or achievement test. For example, measures of high school quality have been developed by Bloom and Peters (1961) and Loeb and Mueller (1970). Tests to measure attitudes and study habits have been developed by Holtzman, Brown, and Farquhar (1954), and socio-economic variables have been added by Barger and Hall (1965).

The purpose of this study was to evaluate the predictive usefulness of several variables which measure some aspects of high school quality, motivation, study time, and socio-economic status. The effectiveness of these variables in forecasting academic achievement is tested in a simple linear regression framework.

*Methodology*

The data for the study were obtained by administering a questionnaire to a sample of students in the Principles of Economics classes in 1973 at the University of Illinois at Chicago Circle (UICC). Table 1 lists the variables along with their means and standard deviations. All data pertain to the academic quarter just prior to the time at which the questionnaire was administered. The sample consisted of unmarried white students, except for four married and two black students. Deletion of these individuals from the statistical analysis to be presented does not alter the conclusions. Students on the GI Bill were excluded because of their unusually high ages and levels of outside financial support. After results with a smaller sample had been obtained, the sample was expanded to its present size of 152 students.

A brief discussion of the hypotheses associated with each of the important additional variables follows. Type of high school attended (central city or suburban and public or parochial) was used as a simple proxy measure of high school quality. Most students at UICC have worked for compensation. It was hypothesized that the greater the number of hours of outside work the lower would be earned grade point average (GPA) because study time per credit hour would be reduced and because students who worked more might have a weaker motivation for success in college. It was also hypothesized that scholarship support should increase grades because students who obtained scholarships exhibited some motivation for success in college. Finally, it was hypothesized that credit hours taken might be positively or negatively associated with GPA because students who took more credit hours might show more motivation but also probably studied less per credit hour.

Grade point average (GPA) for the previous academic quarter was measured on a scale in which  $A = 500$ ,  $B = 400$ ,  $C = 300$ ,  $D = 200$ , and  $E = 100$ .

*Results*

The correlation matrix is presented in Table 1, and the multiple regression results are shown in Table 2. The row in Table 2 labeled regression No. 1 shows the result of regressing GPA on the American College Testing Program (ACT) Composite Score and percentile rank in high school class. The multiple correlation coefficient for this regression was .295, somewhat lower than the .50 found by Loeb and Mueller (1970) in a similar regression for a sample of UICC students. In regression analysis No. 2 a set of dummy variables indicating type of

TABLE 1  
Variable Definitions, Means, Standard Deviations, and Correlations  
(*N* = 152)

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	Mean	Std. Dev.
1. Grade Point Average	—												361.6	64.33
2. ACT Composite Score	26	—											23.54	3.61
3. Percentile Rank, HS Class	18	18	—										78.45	18.48
4. City Parochial HS	-07	-06	21	—									.303	
5. Suburban Public HS	06	14	-31	-39	—								.263	
6. Suburban Parochial HS	18	09	-02	-24	-22	—							.118	
7. HS Not In Chicago Area	-01	-16	03	-08	-07	-04	—						.013	
8. Sex (1 = female, 0 = male)	19	-04	00	25	-18	-13	-06	—					.197	
9. Scholarships (\$/quarter)	-06	-06	05	04	-19	05	-06	-01	—				97.30	201.60
10. Hours Worked Per Week	-21	02	-12	-02	04	06	-11	-15	-10	—			16.25	11.48
11. Credit Hours Taken	08	08	-03	07	01	-08	01	16	03	-21	—		14.22	5.75
12. Non-city HS	18	14	-28	-53	74	45	14	-27	-26	05	-04	—	.394	

TABLE 2  
*Regression Analysis of Grade Point Averages*  
*(Prediction Variables Included are Numbered in Table 1)*

Identification of the Regression Analysis	Regression Equations and Corresponding Multiple Correlation Coefficients			
No. 1	GPA = 274.7 + 4.169 (Var. 2) + .496 (Var. 3)			Multiple R = .295
	(7.74)	(2.93)***	(1.78)*	
No. 2	GPA = 285.9 + 3.395 (Var. 2) + .691 (Var. 3)			+ 1.395 (Var. 4) + 22.03 (Var. 5)
	(7.74)	(2.31)**	(2.34)**	(.11) (1.58)
		+ 41.94 (Var. 6) + 18.58 (Var. 7)		Multiple R = .365
		(2.43)**	(.41)	
No. 3	GPA = 258.9 + 3.219 (Var. 2) + .783 (Var. 3)			+ 41.12 (Var. 12) + 38.35 (Var. 8)
	(7.47)	(2.42)**	(1.89)***	(3.91)*** (3.13)***
		+ .044 (Var. 9) - .665 (Var. 10) + 1.924 (Var. 11)		Multiple R = .518
		(1.88)*	(1.59)	(2.31)**

Note.—The *t* statistics are in parentheses below the coefficient estimate. The designations \*, \*\*, and \*\*\* indicate significance at the .10, .05, and .01 levels, respectively, for a two-tail test.

high school attended was added to regression analysis No. 1. The results indicate that graduating from a suburban parochial school added significantly to grades (.42 of a letter grade), and that the suburban public school background might add to GPA. The increase in explanatory power over regression No. 1 was highly significant ( $F = 118.5$ ). It should also be noted that the coefficient of high school rank was estimated with more precision in regression analysis No. 2 than in regression analysis No. 1. In regression analysis No. 3 the socio-economic variables were included and the high school type variable was used as a dummy variable which indicated that the student graduated from a high school not in the city of Chicago. The results revealed that female students did achieve significantly higher grades than male students did (.38 of a letter grade), students who took more credit hours made slightly higher grades than those who took fewer credit hours, students who worked might receive slightly lower grades than those who did not work, and that students with scholarship support might earn higher grades than might those without such support. The hypothesis that the relationship between GPA and hours of work was linear could not be rejected ( $F = 1.90$  with 1 and 134 degrees of freedom). Additional variables were tested in regression analysis not reported here. These variables, which included the student's age, year in school, number of siblings and the parents' income and educational attainment, were all statistically nonsignificant.

### Interpretation

Except for sex, none of the standard measures of socio-economic status made a significant contribution to the prediction of college



success. This failure might stem from the fact that a biased sample of college students with high status parents had been observed. Because UICC is an inexpensive state university, many parents of high socioeconomic status may prefer to send their more highly motivated children to more expensive and higher-status colleges. In addition, female students at UICC may earn higher grades because they exhibit stronger motivation or because they are self-selected according to some ability characteristic which has not been measured. Further validation of the results with new samples is warranted.

### REFERENCES

- Barger, B. and Hall, E. The interaction of ability levels and socioeconomic variables in the prediction of college dropouts and grade achievement. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1965, 25, 501-508.
- Bloom, B. S. and Peters, F. R. *The use of academic prediction scales for counseling and selecting college entrants*. New York: Free Press of Glencoe, 1961.
- Holtzman, W., Brown, W., and Farquhar, W. The survey of study habits and attitudes: A new instrument for the prediction of academic success. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1954, 14, 726-732.
- Loeb, J. W. and Mueller, D. J. The use of a scale of high schools in predicting college grades. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1970, 30, 381-386.



## RELATIONSHIPS OF SELECTED NONACADEMIC AND ACADEMIC VARIABLES TO THE GRADE POINT AVERAGE OF BLACK STUDENTS<sup>1</sup>

SHIH-SUNG WEN AND ROSE E. MCCOY

Jackson State University

An analysis of correlations of the data mainly from 164 male and 202 female black undergraduate students indicated that (a) a weighted set of measures of manifest needs (Edwards Personal Preference Schedule) correlated significantly with the grade point average (GPA) for the males ( $R = .53$ ,  $df = 15/148$ ,  $F = 3.84$ ,  $p < .001$ ) but not for the females ( $R = .30$ ,  $df = 15/186$ ,  $F = 1.20$ , ns), (b) a weighted composite of measures of personal problems (Mooney Problem Check List) correlated significantly with the GPA for both the males ( $R = .47$ ,  $df = 11/152$ ,  $F = 3.86$ ,  $p < .001$ ) and the females ( $R = .36$ ,  $df = 11/190$ ,  $F = 2.48$ ,  $p < .005$ ), (c) manifest anxiety (Taylor Manifest Anxiety Scale) correlated significantly with the GPA for the male students only ( $r = -.22$ ,  $df = 162$ ,  $p < .005$ ), and (d) the scholastic aptitudes (American College Testing) correlated significantly with the GPA for both male ( $R = .48$ ,  $df = 5/108$ ,  $F = 6.40$ ,  $p < .001$ ) and female students ( $R = .50$ ,  $df = 5/139$ ,  $F = 9.34$ ,  $p < .001$ ).

STUDIES have impressively demonstrated that certain variables are associated with higher or lower academic achievement than would have been predicted from intelligence tests alone. There is considerable evidence that students of high intellectual ability sometimes fail or drop out of college; thus, a measure of intelligence is not a sufficient forecaster of college success.

Behavior analysis of the college experience indicates that non-academic variables are highly relevant to the quality and rate of academic performance. Quantitative indication of the learner's motiva-

<sup>1</sup> The study was supported by the Research and Publication Committee at Jackson State University.

tion, anxiety, personal problems, among others, have been widely accepted as important determinants of college success.

The purpose of this study was to ascertain for each of two samples of 164 male black and 202 female black students in a southern state university the degree to which selected nonacademic and academic characteristics of college students as indicated by self-report scales and standardized measures of scholastic ability were predictive of college success as revealed by grade point average (GPA) earned during the first two quarters of the 1972-73 academic year. Specifically, for each sample the degree of correlation was sought between GPA and (a) each of the 15 measures of manifest needs within the Edwards Personal Preference Schedule (EPPS) (Edwards, 1959); (b) each of the 11 measures of personal problems within the Mooney Problem Check List (MPCL) (Mooney and Gordon, 1950); (c) the single measure of anxiety obtained by the Taylor Manifest Anxiety Scale (TMAS) (Taylor, 1953); and (d) each of five measures of scholastic ability provided by the American College Testing Program (ACT) including English, Mathematics, Social Studies, Natural Sciences, and Composite Score (1968-1971). In addition, multiple correlation coefficients were also determined between the GPA and each of the weighted composites of measures of the EPPS, MPCL, and ACT.

The sample for the study was drawn from the student population of a predominantly black state university. Since less than 5% of the student enrollment is nonblack, and since the state in which the university is located is largely rural, it could be assumed that the majority of the student population has come from relatively depressed environments, and further, that achievement motivation, personal adjustment, and other relevant factors have been unfavorably shaped by the experiences in these environments.

In a study of motives, Brazziel (1964) reported that both lower and middle class black females in the lower-south manifested high needs for achievement. However, Williams and Cole (1969) reported apathy and low morale toward academic achievement in black students. Atchinson (1968) found a significant positive correlation between Manifest Anxiety Scale scores and grade point averages for black college sophomores. The correlation between the ACT assessment and college grades was .59 among freshmen at a southern black state college (Funches, 1965).

### *Method*

Subjects were 164 male and 202 female students in sections of general psychology. Their ages ranged from 18 to 23 with the average about 19.

Over a period of two quarters of the academic year 1972-1973, the EPPS, MPCL, TMAS were administered to students during the regular class periods. The EPPS consists of 15 subscales: Achievement (*Ach*), Deference (*Def*), Order (*Ord*), Exhibition (*Exh*), Autonomy (*Aut*), Affiliation (*Aff*), Intracception (*Int*), Succorance (*Suc*), Dominance (*Dom*), Abasement (*Aba*), Nurturance (*Nur*), Change (*Chg*), Endurance (*End*), Heterosexuality (*Het*), and Aggression (*Agg*). The 11 problem areas of the MPCL are: Health and Physical Development (HPD); Finance, Living Conditions, and Employment (FLE); Social and Recreational Activities (SRA); Social-Psychological Relations (SPR); Personal-Psychological Relations (PPR); Courtship, Sex, and Marriage (CSM); Home and Family (HF); Morals and Religions (MR); Adjustment to School Work (ACW); The Future—Vocational and Educational (FVE); and Curriculum and Teaching Procedure (CTP).

Subjects' scores on the ACT and their cumulative GPA's were obtained from college records. Since some students' ACT scores were not available from the records, the *N* was reduced for the correlation between the ACT and GPA.

The product-moment correlations were calculated between the GPA and each of 15 measures of EPPS, each of 11 measures of MPCL, each of five ACT scores, and the single measure of TMAS. In addition, the multiple correlation analyses (Cooley and Lohnes, 1971) were conducted for the GPA and each weighted set of measures of EPPS, MPCL, and ACT, separately.

### Results

Results of multiple correlation analyses and variables each of which correlated significantly (at or beyond the .05 level) with the GPA are presented in Table 1.

In males, the variables, each of which significantly correlated with the GPA, were 14 measures of EPPS, eight subscales within MPCL, five ACT scores, and the single TMAS scale ( $r = -.22$ ). The multiple correlations between the GPA and each weighted set of measures of EPPS, MPCL, and ACT were all significant at the .001 level.

In females, one of EPPS measures, two of MPCL subscales, and five ACT scores significantly correlated with the GPA. Thus, the significant multiple correlations between the GPA and each weighted set of measures were limited to the MPCL and ACT only.

The results indicated that manifest needs, manifest anxiety, and personal problems associated with academic achievement were stronger for male black than for female black college students.



TABLE 1  
Multiple Correlation Analyses and Citation of Variables Significantly Correlated with the GPA

Sex	Predictors	R	df	F	p	Significant Correlations <sup>a</sup>
Male	EPPS	.53	15/148	3.84	.001	Ach (-.26) Def (-.30) Ord (-.33) Exh (-.31) Aut (-.21) Aff (-.31) Int (-.23) Suc (-.35) Dom (-.22) Aba (-.29) Nur (-.28) Chg (-.26) End (-.18) Agg (-.30)
	MPCL	.47	11/152	3.86	.001	FLE (-.16) SPR (-.23) PPR (-.28) HF (-.26) MR (-.26) ACW (-.33) FVE (-.25) CTP (-.26)
	ACT	.48	5/108	6.40	.001	ENG (-.34) MAT (.36) SS (.38) NS (.26) CS (.44)
Female	EPPS	.30	15/186	1.20	ns	Int (.15)
	MPCL	.36	11/190	2.48	.005	CSM (-.16) ACW (-.20)
	ACT	.50	4/139	9.34	.001	ENG (.42) MAT (.25) SS (.39) NS (.30) CS (.48)

<sup>a</sup> Significant at or beyond the .05 level.

## REFERENCES

- Atchinson, C. O. Relationship between some intellectual and non-intellectual factors of high anxiety and low anxiety Negro college students. *Journal of Negro Education*, 1968, 37, 174-178.
- Brazziel, W. F. Correlates of southern Negro personality. *Journal of Social Issues*, 1964, 20, 45-52.
- Cooley, W. W. and Lohnes, P. R. *Multivariate data analysis*. New York: Wiley, 1971.
- Edwards, A. L. *Edwards Personal Preference Schedule*. New York: Psychological Corporation, 1959.
- Funches, D. A. A correlation between the ACT scores and the grade point averages of freshmen at Jackson State College. *College and University*, 1965, 40, 324-326.
- Mooney, R. L. and Gordon, L. V. *The Mooney Problem Check List*. New York: Psychological Corporation, 1950.
- Taylor, J. A. A personality scale of manifest anxiety. *Journal of Abnormal and Social Psychology*, 1953, 48, 285-290.
- Williams, R. L. and Cole, S. Scholastic aptitude of southern Negro students. *Journal of Negro Education*, 1969, 38, 74-77.



## COMPARATIVE PREDICTION OF FIRST YEAR GRADUATE AND PROFESSIONAL SCHOOL GRADES IN SIX FIELDS<sup>1</sup>

LEONARD L. BAIRD

Educational Testing Service

The validity of predictors of academic performance in six post-graduate fields were compared. The fields included three liberal arts areas and three professional areas: arts and humanities, biological and physical science, social science, law, medicine, and business. The predictors included information about students' backgrounds, self-conceptions, values, nonacademic achievements, and curricular patterns as well as admissions test scores and grades. In most fields, grades were predicted by academic ability and by prior achievement, self-confidence, and previous accomplishment in the field. Background variables predicted grades only in law and arts and humanities. The predictive power of admissions tests varied from field to field.

THERE have been many studies of the prediction of academic performance in graduate and professional school (e.g., Lannholm, 1968; Cliff and Cliff, 1972). Although these studies have produced a great deal of information, as a group they have been limited in three ways: (1) they have concentrated on tests of academic ability as predictors; (2) they have usually been limited to a sample of a single department or school; and (3) they have not had comparison groups in other fields. Based on a follow-up of a national sample of college seniors the present study (a) includes information about students' biographical characteristics, self-conceptions, work values, nonacademic achievements, and curricular patterns as well as admissions test scores and grades as predictors; (b) extends over a wide variety of schools and departments; and (c) compares the validity of predictors for samples

---

<sup>1</sup> This study was supported by the Association of American Medical Colleges, the Graduate Record Examinations Board, and the Law School Admissions Council.

of college seniors in relation to criteria of academic performance in six postgraduate fields. The six fields of post graduate study included graduate work in three liberal arts areas and in three professional areas: arts and humanities, biological and physical science, social science, law school, medical school, and (graduate) business school. The purposes of the study were to compare the validity of predictors of academic performance in various fields and to evaluate the contribution of variables assessing students' backgrounds, educational histories, self-conceptions, and values in the prediction of grades.

### *Method*

#### *Data sources*

The data for this study originated from a follow-up of a national survey of a sample of college seniors who replied to a questionnaire, the College Senior Survey, in the spring of 1971 (Baird, Hartnett, and Clark, 1973). Follow-up information regarding the activities of 7,734 seniors in 94 colleges across the country was obtained in late spring of 1972. A variety of data analyses indicated that the sample was representative, except that minority students were slightly under-represented.

The College Senior Survey covered a great deal of biographical, personal, attitudinal, and educational information about students. Reports of their Admission Test for Graduate Study in Business (ATGSB), Law School Admissions Test (LSAT), Graduate Record Examinations (GRE), and Medical College Admission Test (MCAT) scores were obtained. The follow-up questionnaire ascertained students' educational and vocational activities. The criteria used here were self-reported grades in (2) graduate study in the arts or humanities ( $N = 415$ ); (b) graduate study in the social sciences ( $N = 400$ ); (c) graduate study in the biological and physical sciences ( $N = 525$ ); (d) medical school ( $N = 440$ ); (e) law school ( $N = 450$ ); and (f) graduate business school ( $N = 310$ ). The numbers in parentheses indicate the number of cases in each field with complete data in both the senior and follow-up files. Research summarized by the American College Testing Program (1973) has shown self-reported grades to be highly reliable and valid and highly intercorrelated with school reported grades.

#### *Methods*

By generating a missing data correlation matrix of the College Senior Survey information, the writer was able to identify the charac-



teristics that were most strongly correlated with grades in each postgraduate field. In relation to grades in each field as a criterion, stepwise multiple regression analyses were employed to identify the factors that were most strongly associated with grades in each area of study.

Because the undergraduate grades were from such a wide variety of institutions their predictive power in the analyses was probably limited. Similarly, since students in the follow-up sample in any particular area were attending a wide variety of institutions, the size of the correlations of any variable with graduate or professional school grades was probably considerably lower than the value would have been in many single schools.

### *Results*

In Table 1 the zero-order correlation coefficients of each of the predictor variables with grades earned in each of the six fields of graduate and professional study are set forth. In Table 2 the outcomes of stepwise multiple regression analyses are summarized for each field of postgraduate endeavor. Only those variables which yielded statistically significant contributions *and* increased the multiple correlation by at least .01 are listed in Table 2.

Certain variables had higher zero-order correlations in some fields than others as the entries in Table 1 show. Sex and religious background had little relation to grades in any field. Parental education was related only in arts and humanities and business; family income was related only in arts and humanities and law. Parental and peer encouragement of students' plans for further study was unrelated in all fields. Consideration of graduate or professional school at an early age was most positively related to grades in arts and humanities and in the three professional areas. Grades in all courses constituted a more valid predictor than did grades in major field courses in every area except biological and physical science. Several variables reflecting self-confidence of students in their ability to handle academic work were related to grades in every area, most consistently in law and business. Work values generally had small relations to grades in most areas except science, where interest in working with people was negatively related to grades. Admission test scores were less efficient predictors in every field except law than were grades. Tests predicted most accurately in law and business, least so in medicine. Curricular choices were unrelated to grades in most areas except for senior plans to enter the field in medicine, and except for senior major in the field in social science. Almost all nonacademic achievements, such as being president of the student body were unrelated to grades in any area, with the

TABLE I  
Comparative Predictive Correlates of First Year Graduate and Professional School Grades

	Correlation with Grades					
	Graduate Study in			Professional Study in		
	Arts & Hum	Biol/Phs Sci	Social Sci	Law	Medicine	Business
<b>Background Variables</b>						
Sex (1 = Male, 2 = Female)	.06	-.01	.00	-.04	.03	.06
Race (1 = Black, 2 = White)	.10	.08	.08	.08	.05	.06
Parental level of education	.15	.04	.02	.04	.04	.11
Family income	.09	.03	.06	.11	.03	.04
Father's encouragement	-.03	.05	-.04	.00	-.04	-.05
Mother's encouragement	-.03	.01	-.02	-.03	-.05	-.04
Friend's encouragement	-.04	-.07	.03	.04	-.01	.01
Raised in Jewish religion	.02	.01	.01	.09	.08	.06
Age first thought of advanced study	-.15	-.05	-.05	-.09	-.09	-.10
<b>College Grades</b>						
All courses	.23	.21	.31	.26	.26	.31
Major field courses	.18	.26	.28	.21	.24	.23
<b>Self-Conception</b>						
I would rank among the best in academic ability in my class in college	.13	.15	.19	.17	.12	.22
I have the ability to complete the advanced work needed to become a doctor, lawyer, or university professor	.18	.02	.08	.12	.08	.12
I think I would be able to get mostly A's in a graduate or professional school	.09	.11	.11	.18	.12	.18
Self-rating on writing ability	.16	-.01	.08	.12	.03	.19
Self-rating on scholarship	.15	.19	.20	.17	.14	.20
Self-rating on scientific ability	.02	.15	.04	.07	.15	.13
Self-rating on mathematical ability	.03	.11	.05	.14	.17	.07
<b>Work Values</b>						
More interest in people than things	.02	-.11	.05	-.07	.04	.08
Desire to contribute to knowledge	-.05	.08	.02	.06	.01	-.05
<b>Admissions Test Scores</b>						
GRE-V	.19	.16	.18			
GRE-M	.18	.18	.14			
LSAT				.27		
MCAT					.14	
ATGSB						.28
<b>Career Choices</b>						
Freshman vocational choice in field	.00	.00	.03	.04	.03	-.06
Senior major in field	.03	.05	.13	-.03	-.07	-.03
Senior plan to study in field next fall	.05	.05	.02	-.03	.19	.10
Level of degree aspiration	.06	.04	.08	-.03	.06	.09
<b>College Activities</b>						
Won award in field	.08	.12	.03	.04	.08	.22
Assistantship in science	.01	.02	.05	.02	.12	.13

exceptions of having an award in the field in science and business and holding a scientific assistantship in medicine and business.

The general pattern of results obtained in the zero-order correlations was also obtained in the stepwise multiple regression results shown in Table 2. In most fields, grades were predicted by academic ability and achievement, self-confidence, and previous accomplishment in the field. In medicine, the MCAT did not enter as a predictor. Background variables added to the level of prediction only in law and arts and humanities.

### Discussion

Although the requirements for success in graduate and professional fields seem to have some common elements, each one has its unique

TABLE 2  
*Stepwise Multiple Regression Results*

Group	Variables	Final Weight	R	F
Arts & Humanities	College grades in all courses	.19	.23	23.1
	GRE-Mathematical score	.11	.27	16.1
	Parental level of education	.15	.30	13.3
	Age first thought of advanced study	-.09	.31	11.1
	I have the ability to complete the work needed to become a doctor, lawyer or university professor	.10	.32	9.6
Biological & Physical Science		-.06	.33	8.1
	Self-rating on scholarship	.21	.26	38.0
	College grades in major field courses	.11	.29	24.7
	GRE-Verbal score	.08	.30	17.3
	GRE-Mathematical score	.06	.31	13.5
Social Science	Won award in field	.21	.31	42.4
	College grades in all courses	.12	.33	23.9
	College grades in major field courses	.10	.34	17.4
	Senior major in social science	.08	.35	13.8
	GRE-Mathematical score	.19	.27	35.3
Law	LSAT scores	.18	.33	26.9
	College grades in all courses			
	I think I would be able to get mostly A's in a graduate or professional school	.13	.35	21.1
		.10	.37	17.3
	Family income	.15	.26	35.5
Medicine	College grades in all courses	.14	.30	23.4
	Senior plans to study in field	.14	.31	17.6
	College grades in major field courses	.11	.32	14.3
	Self-rating on mathematical ability	.07	.33	11.9
	Held assistantship in science	-.09	.34	10.4
Business	Self-rating—scholarship	.19	.31	32.8
	College grades in all courses	.20	.37	25.0
	ATGSB scores			
	I think I would be able to get mostly A's in a graduate or professional school	.09	.39	18.4
		.09	.40	14.7
	Self-rating on writing ability	.09	.41	12.2
	Won award in field			

pattern. Academic aptitude and achievement were important in every instance. Confidence of students in their ability was also important in every area, especially law and business. Students' conceptions of their abilities had mostly logical relations to the demands of the field. Strictly biographical information made only a small contribution to prediction in most fields, as did work values. Academic success in some areas, such as science and social science seemed to be related most strongly to previous academic performance in the area, whereas in others, such as law, medicine, and business, students' self-confidence plays a role, and in other areas such as arts and humanities and law, success was related to family background.

### REFERENCES

- American College Testing Program. *Assessing students on the way to college*. Iowa City: Author, 1973.
- Baird, L. L., Hartnett, R. T., and Clark, M. J. *The graduates*. Princeton, N.J.: Educational Testing Service, 1973.
- Cliff, M. M. and Cliff, T. M. Attitudes of medical students toward medical school and their future careers. *Journal of Medical Education*, 1972, 47, 534-38.
- Lannholm, G. V. Review of studies employing GRE scores in predicting success in graduate study, 1952-1967. Special report. Princeton, N.J.: Educational Testing Service, 1968.

## THE MODERATOR EFFECT OF UNDERGRADUATE GRADE POINT AVERAGE ON THE PREDICTION OF SUCCESS IN GRADUATE EDUCATION<sup>1</sup>

ROBERT W. COVERT  
University of Virginia

NORMAN M. CHANSKY  
Temple University

Three hundred and six Masters of Education students at a large urban university were divided into six subgroups according to sex and to each of three levels of undergraduate grade point average. Correlation coefficients between graduate grade point average and each of three predictor variables, consisting of Graduate Record Examinations—Verbal score, Graduate Record Examination—Quantitative score, and undergraduate grade point average, were calculated for each of the six subgroups. Results showed differential predictability across the different subgroups.

THE selection of candidates for graduate education programs for many years has depended upon linear regressions. The predictors in these equations have included undergraduate grade point average (UGPA), as well as scores on the Miller Analogy Test (MAT) or Graduate Record Examinations Aptitude Test (Verbal and Quantitative Scores—GREV and GREQ), whereas the criterion has generally been the graduate grade point average (GGPA). The ability of these predictors to differentiate between graduate students varies from institution to institution. In general, however, most prediction studies of graduate success have provided administrators with only one or two validity coefficients for use in the selection of candidates. These coefficients were then assumed to be equally valid for the total group.

---

<sup>1</sup> This research was sponsored by the Internal Research Center of Temple University.



It was the purpose of this research to explore the possibility of differential predictability of candidates on the basis of different levels of UGPA and sex.

### *Method*

#### *Subjects*

The sample of 306 subjects for this investigation included all students who had been accepted into the Master of Education Program at a large urban university from September, 1967 to September, 1968, and for whom GREV, GREQ, UGPA, and GGPA data were complete. These students had either completed or terminated their programs at the time of this investigation.

#### *Procedure*

The total group was divided into thirds on the basis of standing on UGPA. The lowest third included undergraduates with UGPA's of less than 2.5; the moderate group contained students with UGPA's between 2.5 and 2.9; whereas the upper third consisted of persons with UGPA's of greater than 2.9. Multiple correlation coefficients using UGPA, GREQ, and GREV as predictors were then separately calculated for the total group as well as for each of the subgroups which were subsequently divided according to sex.

### *Results*

The means and standard deviations for the total group are included in Table 1.

For the total group, the multiple correlation between the composite of three predictors including UGPA, GREV, and GREQ and the criterion was .29. The correlation between UGPA and GGPA alone was .24.

TABLE 1  
*Means and Standard Deviations for Students Completing or Terminating Masters Programs in Education (N = 306)*

Variables	Mean	Standard Deviation
GREV	533	89
GREQ	496	97
UGPA	2.79	.42
GGPA	3.36	.35

Table 2 illustrates the differential predictability of the total group divided according to thirds on the basis of standing on the UGPA and by sex.

Table 2 shows that when predicting males and females simultaneously, no significant relationship existed between any one of the predictors and GGPA for students whose UGPA was below 2.9. On the other hand, a significant bivariate correlation of .35 existed for those students with UGPA's exceeding 2.9. The table also shows that in the low grade point average group, females could be predicted significantly ( $r = .33$ ); in the moderate grade point range, males could be predicted significantly ( $r = .44$ ); and in the high grade point range, only females could be predicted significantly ( $R = .45$ ).

### Discussion

The findings show that differential predictability was achieved by dividing the students according to sex and UGPA. Important is the fact that those students in the lowest third of UGPA were the least predictable, whereas those in the highest third were the most predictable. Furthermore, the data show that females could be significantly predicted in both the highest and lowest thirds of UGPA, and that males could only be significantly predicted within the moderate range

TABLE 2  
*Correlations, Means, and Standard Deviations for Masters of Education Students Grouped According to Levels of Grade Point Average and Sex*

Group	N	$\bar{X}$	SD	r or R <sup>1</sup>	Predictor(s)
Low Grade Point					
Average (Less than 2.5)					
Female	46	3.33	.26	.33*	GREV
Male	64	3.25	.40	.09	GREV
Total	110	3.29	.35	.15	GREV
Moderate Grade Point					
Average (2.5-2.9)					
Female	66	3.38	.33	.10	UGPA
Male	37	3.34	.31	.44**	GREV
Total	103	3.36	.32	.19	GREV
High Grade Point					
Average (Greater than 2.9)					
Female	71	3.58	.38	.45**	UGPA, GREQ
Male	22	3.34	.29	.14	GREQ
Total	93	3.44	.35	.35**	GREQ

Note.—It should be noted that although multiple correlations were calculated at each level, only in the one case of females in the high UGPA subgroup was the weighted composite of two variables significantly related to the criterion; therefore, the correlation coefficient for females in the high UGPA subgroup represents the only multiple correlation in the table.

of UGPA. These findings should be viewed with caution because of the negative skewness of the scores in the criterion measure. This circumstance might mean that students of low ability and low achievement, when given a chance to complete a master's degree program, might achieve at as high a level as those with greater ability. In looking at Table 2, one can see that although students with high UGPA's had slightly higher average GGPA's, all three groups had GGPA's substantially above the B (3.00) level. One explanation of these high GGPA's might be that little discrimination was made with regard to the quality of performance to which grades had been assigned.

These results have implications for those selecting prospective graduate students in education. First, they suggest that elimination of candidates on the basis of one validity coefficient seems unjustified, since both sex and UGPA had a moderator effect. Second, and perhaps most important, is the fact that the use of the three predictors as the only selection device for candidates would be questionable since, at best, these predictors were accounting for no more than 20% of the variance in the criterion measure of success.

## PREDICTIVE VALIDITY OF SVIB PHARMACIST SCALES<sup>1</sup>

RICHARD W. JOHNSON AND KENNETH W. KIRK  
University of Wisconsin-Madison

RICHARD A. OHVALL  
Ferris State College

The predictive validity of the old and new men's Pharmacist scales and the new women's Pharmacist scale on the Strong Vocational Interest Blank was investigated for 279 male and 76 female pharmacy students. Each of the three scales significantly differentiated between graduates and nongraduates from the pharmacy program. Separate sex norms appeared to be necessary for the two men's scales, but not for the new women's scale. The women's scale most accurately identified the pharmacy majors and produced the highest correlations with the criterion for both sexes.

THE Pharmacist Occupational scale for the Strong Vocational Interest Blank for Men (SVIB-M) recently was revised by Campbell (1971). A Pharmacist scale for the SVIB for Women (SVIB-W) also has been constructed within the past year (Kirk, Johnson, and Ohvall, 1974). The women's single scale primarily measures interest in scientific activities whereas the men's two scales mainly reflect business interests.

Abridged versions of the new men's and women's Pharmacist scales appear on the new Strong-Campbell Interest Inventory (Campbell, 1974). No predictive validity studies for either one of the new men's or women's Pharmacist scales have been reported in the literature.

The primary purpose of this study was to estimate the validity of the Pharmacist scales in predicting graduation from a school of pharmacy.

<sup>1</sup> This study was supported in part by a grant from the Graduate School Research Committee, University of Wisconsin—Madison.

### *Method*

#### *Subjects*

The first-year pharmacy students (college juniors) at the University of Wisconsin-Madison for three successive years (1968-1970) served as subjects. Approximately 90% (279 of 312 males; 76 of 83 females) of the students in the three classes completed the SVIB-M sometime during their first year in the School of Pharmacy.

#### *Predictors*

The SVIB-M (Form T399) was administered to both male and female students in order to obtain comparable data for all students. The original men's scale (Schwebel, 1951), the revised men's scale (Campbell, 1971), and a modified version of the new women's scale were used as predictors. The modified women's scale is based on 60 items found on both the SVIB-M and SVIB-W that differentiated between the interests of 431 women pharmacists and women-in-general by at least 14 percentage points.<sup>2</sup> The SVIB-M version of the women's Pharmacist scale was highly correlated ( $r = .94$ ) with the SVIB-W version for an independent sample of 215 women pharmacists tested by Kirk, Johnson, and Ohvall (1974).

#### *Criterion*

Graduation status at the end of the three year period that was required to complete the pharmacy curriculum was used as the criterion. One-sixth of the male students (48 of 279) and one-fifth of the female students (14 of 76) were not graduated with their class.

#### *Statistical analysis*

The scores on each of the three Pharmacist scales were analyzed by means of two-way analysis of variance (sex  $\times$  graduation status) for unequal cells. The strength of the relationship between the interest scores and graduation status was determined by means of a point biserial correlation. Multiple regression analysis also was undertaken to determine whether the three Pharmacist scales in combination could predict graduation status more effectively than could any one of the individual scales.

<sup>2</sup> A list of the 60 items and the item weights may be obtained from the first author.



*Results and Discussion*

All three of the Pharmacist scales significantly differentiated ( $p < .05$ ) between graduates and nongraduates as shown in Table 1. The magnitude of the difference was approximately five standard score points for both the new men's and women's scales (see Table 2).

To determine whether the relationship between the Pharmacist scales and graduation status was contaminated by an academic ability factor, the Pharmacist scales were correlated with the College Qualification Tests (Verbal, Numerical, Science, Social Studies, Total) for a subsample of 84 of the students who had completed these tests. None of the correlations between the Pharmacist scales and the academic tests was statistically significant ( $p > .05$ ). The effectiveness of the Pharmacist scales in predicting graduation status could not be attributed to the academic ability of the students.

The strength of the relationship between the interest scores and graduation status for each sex was determined by the point biserial correlation coefficients reported in the last column of Table 2. All the correlation coefficients were relatively low. The highest coefficient of .28 accounted for only 8% of the variance in the criterion. Although the interest scores should be helpful in discussing educational and vocational plans with prospective students, they should not be used as a basis for decision-making without additional data.

Neither one of the two multiple correlation coefficients between the weighted composites of the scales and graduation status for either sample of male or female students was significantly greater than was the highest zero-order correlation. The women's scale yielded the highest correlations with the criterion for both samples of men and women. The men's scales did not add significantly to the prediction based solely on the women's scale.

The sexes differed significantly ( $p < .05$ ) in their scores on both of

TABLE 1  
*Two-way Analysis of Variance of Scores on SVIB-M Pharmacist Scales*

Source of Variance	df	Original men's scale	F ratios Revised men's scale	Women's scale
Sex (A)	1	5.24*	14.67***	3.44
Graduation status (B)	1	5.91*	8.51**	12.99***
A $\times$ B	1	.37	.96	.51
Error	351			

\*  $p < .05$ .

\*\*  $p < .01$ .

\*\*\*  $p < .001$ .

TABLE 2  
*Comparison of Mean Standard Scores on SVIB Pharmacist Scales Based on Sex and Graduation Status*

Sex	Graduates		Nongraduates		Total		$r_{pb}$
	Mean	SD	Mean	SD	Mean	SD	
Original men's scale							
Male	38.9	8.9	36.1	10.7	38.4	9.3	.11
Female	36.3	9.1	31.6	10.4	35.4	9.5	.19
Total	38.4	9.0	35.1	10.7			
Revised men's scale							
Male	38.7	12.0	34.7	14.0	38.0	12.4	.12*
Female	32.9	10.8	24.9	13.6	31.4	11.7	.27*
Total	37.4	12.0	32.5	14.4			
Women's scale							
Male	48.0	8.3	43.8	11.0	47.3	8.9	.18**
Female	46.3	8.4	40.0	9.0	45.2	8.8	.28*
Total	47.7	8.3	42.9	10.6			

Note.— $n = 231$  male graduates, 48 male nongraduates, 62 female graduates, 14 female nongraduates

\*  $p < .05$ .

\*\*  $p < .01$ .

the men's scales, but not the women's scale. As shown in Table 2, the mean score for the men was more than one-half standard deviation (6.6 standard scores) higher than was the mean score for women on the new men's scale. The male students apparently endorsed more business interests than did the female students. Separate sex norms should be used for the men's scales; however, combined sex norms are permissible for the women's scale.

The mean scores on the women's scale was considerably higher (9 to 14 standard score points) for both sexes than were the mean scores on either one of the men's scales. Assuming that pharmacy students should obtain high scores on the Pharmacist Occupational scales, the women's scale more effectively identified the students majoring in pharmacy than did either one of the men's scales. The higher scores on the women's scale indicated that both the men and women students had stronger interests in scientific than in business activities.

None of the interactions between sex and graduation status approached statistical significance. The men's scale did not predict more effectively for men than for women, nor did the women's scale predict more effectively for women than for men.

## REFERENCES

- Campbell, D. P. *Handbook for the Strong Vocational Interest Blank*. Stanford, Calif.: Stanford University Press, 1971.

- Campbell, D. P. *Manual for the Strong-Campbell Interest Inventory*. Stanford, Calif.: Stanford University Press, 1974.
- Kirk, K. W., Johnson, R. W., and Ohvall, R. A. Interests of women pharmacists. *Vocational Guidance Quarterly*, 1974, 22, 200-208.
- Schwebel, M. *The interests of pharmacists*. New York: Columbia University Press, 1951.



## USE OF SELECTED FACTORS AS PREDICTORS OF SUCCESS IN COMPLETING A SECONDARY TEACHER PREPARATION PROGRAM

FRANK P. BELCASTRO

Merrimack College  
North Andover, Massachusetts

A multiple regression analysis was performed on all EPPS and SVIB scales to determine those scales which could be used to predict success in a secondary teacher preparation program for 207 females and 88 males. Significant discriminant function equations for fourteen male predictor variables and for eight female predictor variables were obtained. Thus, for both males and females it was possible to discriminate between those who had completed the teacher preparation program and those who had not; for applicants it was possible to classify them as likely or unlikely to complete the teacher preparation program. Cross-validation studies, however, with larger samples would be needed to establish generalizable equations that would permit realization of a comparatively high degree of accuracy of classification of applicant members in new samples.

A major problem of teacher preparation institutions is the selection, from among prospective candidates, of those most likely to be successful in completing a teacher preparation program.

A review of the literature reveals that an impressive number of factors have been investigated in the hope of producing a device which would identify successful teachers in advance of training. These factors or variables were largely achievement tests or were measures of student characteristics (Ayers and Rohr, 1974; Chabassol, 1968; Cook, 1964; Darrow, 1962; Michael, Jones, Gettinger, Hodges, Kolesnik, and Seppala, 1961; Robinson, 1962). This study examined Edwards Personal Preference Schedule (EPPS) and Strong Vocational Interest Blank (SVIB) scales as predictors of success.



### *The Problem*

Normally, application for admittance into the secondary teacher preparation program at Merrimack College, North Andover, Massachusetts is made at the end of the sophomore year. Required for admittance are (a) 2.0 (out of 4.0) overall grade point average, (b) 2.0 grade point average in a major, (c) record of scores on file from a group administration of the EPPS and SVIB measures, and (d) an interview. After verification that the applicant has met the grade point requirements and after exploration of the motivations of the applicant behind his vocational and educational aspirations, philosophy of education, and previous experience with youth, the admittance decision is made mutually by the student and an education department faculty member at the end of the interview. It would be an aid to both the applicant and the education faculty member in arriving at the admittance decision if some prediction could be made from selected EPPS and SVIB scales as to the applicant's successful completion of the teacher preparation program.

The purpose of this study was to determine the usefulness of selected EPPS and SVIB scales as predictors of success in completing a secondary teacher preparation program.

### *Procedure*

Subjects selected for the study were four graduating classes of seniors ( $N=295$ ) who had been admitted to the teacher preparation program two years previously after meeting all requirements for admission to the program.

Part of these graduating seniors ( $N=211$ ) successfully completed the teacher training program including a successful student teaching experience; for a variety of reasons the other graduating seniors ( $N=84$ ) transferred out of the program during the two year period and thus did not complete the program. They also did not have student teaching experience.

The dependent or criterion variable used in this study was defined in terms of whether the graduating senior did or did not complete the teacher training program. Using this criterion of completion versus noncompletion, the investigator endeavored to determine whether the predictor variables could discriminate between these two groups.

The independent or predictor variables consisted of scores on subscales of the SVIB and the EPPS, overall cumulative point average, IQ, and age. All of these predictor variables (51 for the females, 79 for the males) were obtained at the end of the sophomore year at the time of formal entrance into the teacher preparation program.

Scores were analyzed separately for males and females, since the SVIB contained different forms for each sex and since the amount of overlapping between forms was limited. For the completion and non-completion females, the frequencies were 154 and 53, respectively; for the completion and noncompletion males, the numbers were 57 and 31, respectively.

For each of the separate samples of males and females, the null hypothesis to be tested was that there would be no relationship between membership in either one of the two groups (completed or non-completed) and the composite of selected predictor variables.

The Stepwise Multiple Regression (BMD02R) analysis was performed for the males and females on the IBM 360 at the California State University, San Diego.

### *Data Analysis*

To test the general null hypothesis, discriminant equations were computed in a step-wise manner. For the males and females separately, the computer first selected the one "best" independent variable and then selected the "best" of the remaining variables from the pool of predictor variables. This step-wise procedure continued until the entrance of any other predictor variable to the composite did not contribute significantly to the prediction scheme. Only 14 independent variables were chosen for the males and 8 for the females from the pool of predictor variables, since the addition of remaining variables in both cases did not significantly increase predictive effectiveness. These predictor variables are presented in Table 1.

TABLE 1  
*Predictor Variables Which Contributed Significantly to the  
Discriminant Analysis*

DEMOGRAPHIC:	SVIB:
1. Grade Point Average	9. Architect
	10. Physician
	11. Veterinarian
EPPS:	12. Printer
2. Autonomy	13. Industrial Arts Teacher
3. Affiliation	14. Physical Therapist
4. Succorance	15. Artist
5. Nurture	16. Librarian
6. Change	17. English Teacher
7. Heterosexuality	18. Life Insurance Salesman
8. Consistency	19. Femininity-Masculinity Scale
	20. Occupational Level

As inspection of Table 2 reveals that both male and female predictor combinations were significant at the .01 level; hence the null hypothesis was rejected. On the basis of the variables selected for both males and females it would be possible to discriminate between those who had completed the teacher preparation program and those who had not.

Although the multiple correlation coefficients indicated the degree to which it would be possible to predict the dichotomous criterion, they did not yield the kind of information that both the education faculty member and applicant could use directly to arrive at the admittance decision. To facilitate the application of the foregoing analysis, the discriminant function equations for all 14 male predictor variables and all 8 female predictor variables were obtained. The generalized optimizing discriminant equation for the males was:

$$Y_1 = .22285X_1 - .04216X_2 - .01716X_3 - .03183X_4 \\ - .01630X_5 - .01776X_6 + .02169X_7 + .03108X_8 \\ - .01985X_{10} + .01892X_{11} - .02657X_{12} - .00959X_{13} \\ + .01161X_{14} - .01931X_{20} + 2.77798.$$

The corresponding equation for the females was:

$$Y_2 = .21174X_1 + .01762X_7 - .03751X_8 - .01566X_{15} \\ + .01976X_{16} + .00062X_{17} - .00961X_{18} + .00574X_{19} + .30417.$$

To predict the group into which the scores of an applicant would most likely place him, one can substitute into the discriminant equation the actual raw score values for each of the EPPS variables and the report form score values for each of the SVIB variables. Through one's use of the solution to the equation, a male or female applicant could be classified with those likely to complete the teacher preparation pro-

TABLE 2  
Multiple Regression Analyses for Predictor Variables and Criterion Variable  
(Completion vs. Noncompletion) for Males and Females

Predictor Combination	R	R*	SE	F	P
<b>MALES</b>					
$X_1X_2X_3X_4X_5X_6X_7X_8$					
$X_{10}X_{11}X_{12}X_{13}X_{14}X_{20}$	.656	.572	.396	3.941	<.01
$X_1X_7X_8X_{15}X_{16}$					
<b>FEMALES</b>					
$X_{17}X_{18}X_{19}$	.474	.441	.395	7.169	<.01

R\* = shrunken multiple R.

gram if the resulting score was .65 (mean of the male scores) or higher for the male and .74 (mean of the female scores) or higher for the female, and with those unlikely to complete the program if the resulting score was less than the appropriate one of these mean cutoff scores. The greater the deviation from the appropriate mean score, the greater the probability that the applicant's data would more closely parallel those of the particular criterion group.

The efficiency of the equations was measured by obtaining a percentage composed of the ratio of students who were predicted to complete or not to complete the program to students who actually did or did not complete the program. The efficiency of the predictive equations follows: (1) Males—for the total group, the prediction was 78% correct; of those who completed the program, 74% were correctly predicted as successfully completing the program; of those who did not complete the program, 87% were correctly predicted as not completing the program. (2) Females—for the total group, 66% correct; of those who completed the program, 63% correct; of those who did not complete the program, 75% correct.

As a further improvement, the efficiency of the equations was obtained as just explained except that only the top 25% and the bottom 25% of the students were used. This improved efficiency of the predictive equations follows: (1) Males—for the total group, 89% correct; of those who completed the program, 88% correct; for those who did not complete the program, 91% correct. (2) Females—for the total group, 74% correct; of those who completed the program, 67% correct; of those who did not complete the program, 90% correct. Cutoff scores for this group were .8587 for the high and .4365 for the low subgroup of males; .8778 for the high and .6000 for the low subgroups of females.

### *Conclusions*

For both males and females it was possible to discriminate between those who had completed the secondary teacher preparation program and those who had not.

For both males and female applicants, it was possible to classify them as likely or unlikely to complete the teacher preparation program.

For both males and females, applicants could be classified with greater accuracy as unlikely to complete the teacher preparation program than as likely to do so.

Before these prediction equations could be used to assist education faculty members and applicants in arriving at the admittance decision,

it would be necessary to carry out cross-validation studies with large samples to ascertain whether the same predictor variables would be selected and whether the weights assigned to them would remain relatively stable. Since there were so many predictor variables in relation to the number of individuals studied and hence a comparatively small number of degrees of freedom, the larger samples would permit a more nearly realistic estimate of the correlation and an improvement in accuracy of classification when the equations are applied to new samples. Since this study furnished evidence about prediction of success for entering students in only one selective teacher preparation program at only one institution, generalizations at best would be highly tentative.

### REFERENCES

- Ayers, J. B. and Rohr, M. E. Relationship of selected variables to success in a teacher preparation program. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1974, 34, 933-937.
- Chabassol, D. The possession of certain attitudes as predictors of success in practice teaching. *Journal of Educational Research*, 1968, 56, 304-306.
- Cook, D. L. The personal data form as a predictor of success in a teacher education program and entry into teaching. *Journal of Teacher Education*, 1964, 15, 61-66.
- Darrow, H. D. The relationship of certain factors to performance of elementary student teachers with contrasting success records in student teaching. *Teacher College Journal*, 1962, 33, 95-98.
- Michael, W. B., Jones, R. A., Gettinger, T. Jr., Hodges, J. D. Jr., Kolesnik, P. E., and Seppala, J. The prediction of success in selected courses in a teacher training program from scores in achievement tests and from ratings on a scale of directed teaching performance. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1961, 21, 995-999.
- Robinson, W. A validity study of the testing program for the selection of students for teacher education. Unpublished doctoral dissertation, Purdue University, 1962.



## THE PREDICTION OF PERFORMANCE IN AN EDUCATIONAL PSYCHOLOGY MASTER'S DEGREE PROGRAM

ANDREW BEAN  
Temple University

This study examined the predictive validity of the Graduate Record Examinations Aptitude Test, Verbal and Quantitative scores (GREV and GREQ) and undergraduate grade-point average (UGPA). Criterion variables consisted of graduate grade-point average (GGPA), the Master's Comprehensive Examination scores (MCE), and grades in individual required courses. Subjects were 91 students enrolled in the Department of Educational Psychology Master's degree program at a large metropolitan university. GREV correlated .31 with GGPA, but failed to correlate significantly with any other criterion. GREQ correlated .45 and .59 with grades in two research methods courses, but failed to correlate with any other criterion. UGPA was not significantly related to any of the criteria. These somewhat atypical findings stress the need for local validation of graduate admissions measures.

THE predictive validity of measures used for admission to graduate school can vary considerably as a function of the specific criterion used. In many graduate programs, students are required to achieve a minimum grade-point average, to obtain minimum grades in individual courses, and to pass a comprehensive examination in order to be awarded a degree. Thus, the validity of admissions measures needs to be examined using other criteria in addition to the usual graduate grade-point average (GGPA).

Predictive validity studies of the Graduate Record Examinations Aptitude Test, Verbal and Quantitative scores (GREV and GREQ) typically have used GGPA as the single criterion of academic performance. Recent studies conducted within colleges of education have

reported correlations between GREV and GGPA ranging from the .20's to the .30's; correlations between GREQ and GGPA have ranged from low and nonsignificant to as high as .37 (Borg, 1963; Madaus & Walsh, 1965; Payne, Wells, and Clark, 1971).

The predictive validity of undergraduate grade-point average (UGPA) has also been investigated through using GGPA as the criterion. In a representative study, Ayres (1971) found correlations between UGPA and GGPA ranging from .28 to .69 in three different curricular groups within a college of education.

The purpose of this study was to determine the predictive validity of GREV, GREQ, and UGPA for a variety of academic performance criteria in the Department of Educational Psychology Master's degree program at a large metropolitan university. The criteria included grades in individual courses, comprehensive examination performance, and cumulative graduate grade-point average.

### *Variables*

The predictor variables were GREV, GREQ, and UGPA. The criterion variables were as follows: (1) graduate grade-point average (GGPA); (2) Master's Comprehensive Examination score (MCE); and (3) individual required core course grades in Group Processes, Human Development, Learning Theories, Research Procedures, and Survey Research. UGPA, GGPA and individual course grades were computed using the usual four point system (A=4, B=3, C=2, D=1, F=0). MCE scores were obtained from essay responses to questions based on the content of the core courses just specified.

### *Sample*

The subjects were drawn from 91 students enrolled in a General Educational Psychology Master's degree program in a large urban university. These students took the MCE for the first time between Fall 1969 and Spring 1971.

### *Results*

Means and standard deviations for predictor and criterion variables are given in Table 1. Correlations among the predictors and criteria are presented in Table 2. Sample sizes for each correlation are cited in parentheses adjacent to the correlation coefficient.

As indicated in Table 2, GREV correlated .31 with graduate grade-point average; GREQ and UGPA were uncorrelated with GGPA. A

TABLE 1  
Means and Standard Deviations for Predictor and Criterion Variables

	Mean	Standard Deviation	N
<i>Predictor variables:</i>			
GREV	571	83.40	60
GREQ	520	95.30	60
UGPA	2.78	.34	86
<i>Criterion variables:</i>			
GGPA	3.41	.27	91
Group Processes	3.58	.50	66
Human Development	3.33	.53	75
Learning Theories	3.28	.53	75
Research Procedures	2.95	.73	62
Survey Research	3.35	.71	31
MCE	49.60	6.81	91

stepwise regression analysis, undertaken to determine whether some optimal weighted combination of GREV, GREQ, and UGPA would more accurately predict GGPA than would the GREV alone, indicated that combining GREQ and UGPA and GREV failed significantly to increase the predictability of graduate grade-point average. Total GRE score (a simple sum of GREV and GREQ) correlated .25 with GGPA. It should be noted that this coefficient is lower than the .31 correlation coefficient obtained from using GREV alone. Thus, GREQ was not a useful predictor of graduate grade-point average when used alone or in combination with GREV.

GREQ correlated .45 with grades in Research Methods and .59 with grades in Survey Research. The content of these two courses included an introduction to elementary statistical methods. Thus, GREQ was a useful predictor of performance only in courses with a strong quantitative emphasis.

TABLE 2  
Correlations among Predictor and Criterion Variables\*

Criterion Variables	Predictor Variables		
	GREV	GREQ	UGPA
Graduate GPA	.31* <sub>(60)</sub>	.10 <sub>(80)</sub>	.05 <sub>(86)</sub>
Group Processes	.02 <sub>(40)</sub>	-.02 <sub>(40)</sub>	-.10 <sub>(61)</sub>
Human Development	.09 <sub>(48)</sub>	-.04 <sub>(48)</sub>	.05 <sub>(70)</sub>
Learning Theories	.04 <sub>(49)</sub>	.11 <sub>(49)</sub>	-.02 <sub>(72)</sub>
Research Procedures	-.07 <sub>(40)</sub>	.45** <sub>(40)</sub>	-.11 <sub>(80)</sub>
Survey Research	.10 <sub>(24)</sub>	.59** <sub>(24)</sub>	.13 <sub>(29)</sub>
MCE	.19 <sub>(60)</sub>	-.13 <sub>(60)</sub>	.13 <sub>(80)</sub>

\*  $p < .05$ .

\*\*  $p < .01$ .

\* Sample sizes specified in parentheses.

GREV, GREQ, and UGPA were not significantly correlated with MCE score. However, GGPA correlated .59 with MCE scores. As would be expected, a student's performance on the MCE was much more closely related to the grades he received in courses than to measures of his academic aptitude obtained prior to admission.

### *Discussion*

The correlation obtained between GREV and GGPA was typical of those reported by other investigators (Borg, 1963; Madaus and Walsh, 1965; Payne et al., 1971). However, the failure of UGPA to predict any criterion of academic performance was somewhat surprising in light of the findings of other investigators (Ayres, 1971; Payne, et al., 1971). Restricted range in UGPA did not appear to limit predictability, since UGPA scores ranged from 2.0 to 3.6. No evidence of curvilinearity was found.

One possible explanation of the low correlation between UGPA and GGPA is that the Department of Educational Psychology Master's degree students came from an unusually wide variety of undergraduate colleges and curricula. The meaning of a "3.00" average varied depending upon the college and curriculum in which it was earned. Thus, the noncomparability of undergraduate records may have made them worthless as predictors in this case.

Although GREV was a significant predictor of GGPA when used alone, prediction was not improved by forming a weighted composite of GREV and GREQ. GREQ was shown to be a useful predictor of grades in research courses, whereas GREV was not. Thus, the data indicate that high quantitative ability did not compensate for low verbal ability in predicting total graduate grade-point average; similarly, high verbal ability did not compensate for low quantitative ability in predicting performance in research courses. Prediction using a weighted composite, as in multiple regression, rests on such a compensation hypothesis. If performance in both total GGPA and individual research courses is considered important, minimum scores for admission should be set separately for GREV and GREQ, rather than using some weighted composite.

A practice commonly followed in making admission decisions is to set a minimum GRE total aptitude score, instead of using GREV and GREQ separately. For the data in this sample, such a practice actually *reduced* the predictive validity from that obtained in using GREV alone. Thus, the validity of the GRE total aptitude score should be checked empirically, rather than simply assumed.

The lack of predictive validity shown by UGPA in this sample

further emphasizes the necessity for local validation of admissions measures. The fact that a measure have shown predictive validity in a number of settings is no guarantee of validity in a particular location.

### REFERENCES

- Ayres, J. B. Predicting quality point averages in a master's degree program in education. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1971, 31, 491-495.
- Borg, W. R. GRE aptitude scores as predictors of GPA for graduate students in education. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1963, 23, 379-382.
- Madaus, G. F. and Walsh, J. J. Departmental differentials in the predictive validity of the Graduate Record Examination aptitude tests. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1965, 25, 1105-1110.
- Payne, D. A., Wells, R. A., and Clark, R. R. Another contribution to success in graduate school: A search for sex differences and comparison between three degree types. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1971, 31, 497-504.





## THE RELATIONSHIP OF THE WATSON-GLASER CRITICAL THINKING APPRAISAL TO SEX AND FOUR SELECTED PERSONALITY MEASURES FOR A SAMPLE OF DUTCH FIRST-YEAR PSYCHOLOGY STUDENTS

JOH. HOOGSTRAATEN AND H.H.C.M. CHRISTIAANS<sup>1</sup>

University of Amsterdam

In reaction to an earlier publication by Simon and Ward (1974) on the 1952 version of the Watson-Glaser Critical Thinking Appraisal, data are presented on the relationship of the 1964 forms of the same instrument to four selected noncognitive measures for a sample of 190 Dutch psychology students. Except for Subtest 5, Evaluation of Arguments, subtest and total score means were significantly lower for Form ZM than for Form YM. Reliabilities of the Watson-Glaser (W-G) subtests ranged from only .22 to .69. Total score KR-20 reliability estimates, however, were .72 (ZM) and .77 (YM).

No sex differences were found. The correlation between the W-G total scores and those on the extroversion-introversion measure was not significant. Correlations of the W-G measure with other personality characteristics (neuroticism and rigidity) were also close to zero. As for version ZM of the W-G measure the performance was significantly associated with test-defensiveness.

The present research was undertaken to study the concurrent validity of the Watson-Glaser Critical Thinking Appraisal (Watson and Glaser, 1964, Form YM and ZM), relative to five affective measures with Dutch male and female first-year psychology students. To achieve this objective, the results obtained with Dutch students were compared with those found by Simon and Ward (1974) with British students. Simon and Ward studied the relationship between scores on the 1952

---

<sup>1</sup> The writers thank Prof. dr. F. N. Kerlinger for his comments on an earlier draft of this article.

TABLE 1

*Means and Standard Deviations of Scores of Psychology Students on the Watson-Glaser Critical Thinking Appraisal, Forms YM and ZM*

Subtests and Total scale	Maximum Score	Means		SDs		Student <i>t</i>
		YM <i>n</i> = 97	ZM <i>n</i> = 96	YM <i>n</i> = 97	ZM <i>n</i> = 96	
1 Inference	20	12.20	11.20	2.49	2.83	2.595 ( <i>p</i> < .05)
2 Recognition of Assumptions	16	13.62	12.98	1.59	1.52	2.846 ( <i>p</i> < .05)
3 Deduction	24	20.31	18.15	3.19	2.58	5.141 ( <i>p</i> < .001)
4 Interpretation	24	19.94	17.45	2.19	2.91	6.689 ( <i>p</i> < .001)
5 Evaluation of Arguments	15	10.46	11.09	2.35	2.00	-1.996 ( <i>p</i> > .05)
Total	100	76.56	70.31	7.75	9.16	5.091 ( <i>p</i> < .001)

version of the Watson-Glaser and (a) sex and (b) scores on a measure of introversion-extroversion. In the current study, the association of the test with measures of neuroticism, rigidity, and test-defensiveness was also determined. A negative relation between the Critical Thinking ability and each of these three characteristics was hypothesized. Finally, subtest intercorrelations, subtest-total scale correlations, and reliability coefficients (KR-20) of both the YM and ZM form were obtained.

### Procedure

Ninety-seven male and 92 female first-year psychology students of the University of Amsterdam, about two-thirds of all the psychology freshmen, responded as part of their course obligations, either to the YM or the ZM form of the Watson-Glaser measure, which had been translated into Dutch. The allocation of the two forms to the total sample was random. Ninety-three of the male students and 77 of the female, furthermore, completed personality questionnaires on rigidity (Tellegen, 1968), Neuroticism (*N*), Neuroticism Manifested by Somatic Complaints (*NS*), Extroversion-Introversion (*E*), and Test-Taking Attitude (*T*). These four last-mentioned characteristics were as-

TABLE 2

*Reliability (KR-20) Estimates of the Watson-Glaser Forms YM and ZM*

Forms	Subtests					Total
	1	2	3	4	5	
YM	39	40	69	42	58	77
ZM	49	22	44	55	44	72

TABLE 3  
*Subtest-Intercorrelations and Subtest-Total Correlations for Forms  
 YM and ZM of the Watson-Glaser*

Subtests and Total scale	1		2		3		4		5	
	YM	ZM	YM	ZM	YM	ZM	YM	ZM	YM	ZM
Reference	—	—	—	—	—	—	—	—	—	—
Recognition of Assumptions	.24	.03	—	—	—	—	—	—	—	—
Induction	.31	.27	.30	.22	—	—	—	—	—	—
Interpretation	.21	.31	.20	.32	.24	.22	—	—	—	—
Evaluation of Arguments	.33	.11	.01	.07	.17	.22	.15	.27	—	—
Total	.66	.63	.47	.43	.55	.51	.62	.58	.46	.52

vised by the Amsterdamse Biografische Vragenlijst, the Dutch version of Eysenck's M.P.I. (Wilde, 1963)

### Results

The results of univariate *t* tests indicate that for the sample studied Form YM was less difficult than was Form ZM. The data are given in Table 1, the means, standard deviations, and *t* ratios of the two samples (Forms YM and ZM) of the five subtests as well as the total scores. The difference between the means of the total scores was significant at the .001 level. The difference between the means of the five subtests were also significant. The only exception to the general pattern that Form ZM was more difficult than was Form YM was subtest 5 (Evaluation of Arguments). The difference however was not substantial (.63), though significant ( $p < .05$ ). On the whole the subtest reliabilities shown in Table 2, were low in that they ranged from .22 to .69. For the total scales, however, higher KR 20 values of .77 (YM) and .72 (ZM) were obtained. As for subtest intercorrelations and subtest-total scale correlations, which are given in Table 3 it is clear that the correlations were low.

Subtest-total scale correlation coefficients were more substantial

TABLE 4  
*Correlation Coefficients between the Watson-Glaser Test Scores and Five  
 Personality Characteristics of Male and Female Psychology Students*

Personality Characteristics	Form YM		Form ZM	
	Male ( $n = 46$ )	Female ( $n = 33$ )	Male ( $n = 45$ )	Female ( $n = 62$ )
Neuroticism (N)	.16	.11	.15	.20
Neuroticism (NS)	.07	.02	.01	.09
Extraversion (E)	.07	.08	.05	.12
Test-Defensiveness (T)	.13	.04	.26	.30

than were the intercorrelations of the subtests, as the coefficients varied from .43 to .78, though the obtained values were overestimates because the subtests are, of course, parts of the total scales.

Correlations between the Watson-Glaser measure and each of the four personality characteristics are presented in Table 4. With one exception, the measure of test-defensiveness on the ZM form, no significant values were obtained. Moreover, the coefficients of correlation of the total scales with rigidity were low and also not significant:  $r = .10$  (YM)  $r = -.01$  (ZM). Finally, the mean differences between the total scores of male and female subjects did not reach statistical significance. Men obtained means of 76.73 (YM) and 71.53 (ZM); women 76.51 (YM) and 70.26 (ZM).

### *Discussion*

Supporting the data reported by Simon and Ward with British students, the results with the Dutch students suggest that sex is not a significant factor in performance on the Watson-Glaser Critical Thinking Appraisal, Forms YM and ZM. The data also indicate that performance on these equivalent critical thinking instruments is probably independent of two personality characteristics, neuroticism and rigidity. Furthermore the results confirm the conclusion reached by Simon and Ward on the 1952 version that there is no relation between the Watson-Glaser measure and the Extroversion-Introversion scale. It was shown that there was a significant negative relation between the critical thinking performance and test-defensiveness, though this result held only for Form ZM and the correlation did account for no more than about 10% of the variance.

As can be seen from Table 1 Form ZM seems to be more difficult than Form YM. This result upholds the position taken in the manual of the Watson-Glaser Critical Thinking Appraisal which states that separate percentile norms hold for Forms YM and ZM.

The manual furthermore reports low subtest intercorrelation coefficients. The data support the idea that these subtests measure relatively distinct abilities. This result can be at least partly explained, however, by the moderately low reliability coefficients of the five subtests of both forms.

Though overestimated the total scales reach more nearly acceptable reliability levels than the subtests do.

### REFERENCES

- Simon, A. and Ward, L. O. The performance on the Watson-Glaser Critical Thinking Appraisal of university students classified ac-



ording to sex, type of course pursued, and personality score category. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1974, 34, 957-960.

Tellegen, B. *Over rigiditeit*. Zaltbommel, 1968.

Watson, G. and Glaser, E. M. Watson-Glaser Critical Thinking Appraisal. Manual. New York: Harcourt, Brace and World, 1964.

Wilde, G. J. S. *Neurotische labiliteit gemeten volgens de vragenlijstmethode*. Amsterdam, van Rossen N.V., 1963.



## CONVERGENT AND DISCRIMINANT VALIDITIES OF TWO SETS OF MEASURES OF SPATIAL ORIENTATION AND VISUALIZATION

LEWIS PRICE AND JOHN ELIOT

University of Maryland

Responses from high school sophomores to the Eliot-Price and the Guilford-Zimmerman tests of spatial orientation and visualization were evaluated in terms of the multitrait-multimethod matrix technique. The Eliot-Price tests were found to meet all criteria for convergent and discriminant validity. However, the Guilford-Zimmerman tests did not meet one of the criteria for discriminant validity. The Eliot-Price tests appeared to be more nearly precise measures of spatial orientation and visualization.

BORICH and Bauman (1972) examined the convergent and discriminant validities of the French (1962) and the Guilford-Zimmerman (1956) tests of spatial orientation and visualization. Using the multitrait-multimethod matrix technique (Campbell and Fiske, 1959), they found that the variance attributed to method exceeded the variance attributed to traits, and concluded that although there was evidence for convergent validity, there was little evidence for discriminant validity (Table 1).

### *Purpose and Procedure*

The purpose of this research was to make a similar comparison of the Eliot-Price and the Guilford-Zimmerman tests of spatial orientation and visualization. Price (1974) described the Eliot-Price tests in detail, and also showed that the pair of tests correlated highly with other putative measures of the same abilities. In the present study, the two sets of tests were given to 39 randomly selected high school

TABLE 1  
*Multitrait-Multimethod Matrix for French and  
 Guilford-Zimmerman SR-O and Vz Tests*

		Method			
		Guilford-Zimmerman		French	
	Trait	SR-O	Vz	SR-O	Vz
Guilford-Zimmerman	SR-O	(.88) <sup>a</sup>			
	Vz	.67	(.93) <sup>b</sup>		
French	SR-O	.48	.53	(.66) <sup>c</sup>	
	Vz	.34	.44	.55	(.51)

Note.—All correlation coefficients were significant beyond the .05 level.

<sup>a</sup> Alternate forms reliability reported by Guilford and Zimmerman (1956).

<sup>b</sup> Kuder-Richardson 21 reliability reported by Guilford-Zimmerman (1956).

<sup>c</sup> Alternate forms reliability reported by the authors.

sophomores, and their responses were evaluated in terms of the multitrait-multimethod matrix technique (Table 2).

### Results and Discussion

The values in the diagonal, representing the convergent validity data from this study, were higher than were those obtained by Borich and Bauman (.60 and .62 compared with .48 and .44 respectively). The low reliabilities of the French tests might account for the low convergent validity coefficients in Table 1. Indeed, the French tests might not be

TABLE 2  
*Multitrait-Multimethod Matrix for Eliot-Price and Guilford-Zimmerman SR-O  
 and Vz Tests*

		Method			
		Guilford-Zimmerman		Eliot-Price	
	Trait	SR-O	Vz	SR-O	Vz
Guilford-Zimmerman	SR-O	(.88) <sup>a</sup>			
	Vz	.71	(.93) <sup>b</sup>		
Eliot-Price	SR-O	.60	.51	(.80) <sup>c</sup>	
	Vz	.52	.62	.52	(.93) <sup>d</sup>

Note.—All correlation coefficients were significant beyond the .05 level.

<sup>a</sup> Alternate forms reliability reported by Guilford-Zimmerman (1956).

<sup>b</sup> Kuder-Richardson 21 reliability reported by Guilford-Zimmerman (1956).

<sup>c</sup> Coefficient alpha reported by authors.

<sup>d</sup> Coefficient alpha reported by Price (1974).

as stable measures of spatial orientation and visualization as those of Guilford-Zimmerman and Eliot-Price.

The correlations (.60 and .62) between independent methods (different authored tests) measuring the same intended traits of SR-O and Vz exceeded the correlations (of .51 and .52) between either one of them and other traits not having the method (same authored tests) in common. Thus, the criterion of discriminant validity was met satisfactorily. Similarly, the correlations (.60 and .62) between independent methods (different authored tests) measuring the same intended trait exceeded the correlation of .52 between different traits of SR-O and Vz which employed the same method or same authored test by Eliot and Price. However, in the instance of the Guilford-Zimmerman tests, the criterion of discriminant validity was *not* met, as the correlation of .71 between the different traits of SR-O and Vz *exceeded* the correlation of .60 and .62 between the different authored tests on the respective same traits of SR-O and Vz. In this latter case, the criterion of discriminant validity was more satisfactorily met for the Eliot-Price tests than for the Guilford-Zimmerman tests. It appears that the Guilford-Zimmerman tests might have another factor in common which is indicated by the .71 correlation. The ability to understand complex verbal instructions could conceivably be that other common factor.

The Eliot-Price tests appeared to be more nearly precise measures of spatial orientation and visualization in that they met all criteria for convergent and discriminant validity.

## REFERENCES

- Borich, G. D. and Bauman, P. M. Convergent and discriminant validation of the French and Guilford-Zimmerman spatial orientation and spatial visualization factors. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, 1972, 32, 1029-1033.
- Campbell, D. T. and Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, 56, 81-105.
- French, J. W., Ekstrom, R. B., and Price, L. A. Kit of reference tests for cognitive factors. Princeton, New Jersey: ETS, 1962.
- Guilford, J. P. and Zimmerman, W. S. Guilford-Zimmerman aptitude survey. Beverly Hills, Calif.: Sheridan, 1956.
- Price, L. C. The validation of a Spatial Visualization Test using a Piagetian task. Unpublished Master's Thesis, University of Maryland, 1974.





## FACTORIAL DIMENSIONS OF THE JESNESS INVENTORY WITH BLACK DELINQUENTS

ROGER WOODBURY

Wilson County Technical Institute  
Wilson, North Carolina

JAMES SHURLING

North Carolina State University

The study identified the personality dimensions in the Jesness Inventory (JI) among black male delinquents. A random sample of 250 black male delinquents was administered the JI. A principal components factor analysis with a varimax rotation identified three factors: (1) Self-Estrangement, (2) Social Isolation, and (3) Immaturity. The proportions of total common-factor variance accounted for by the three factors were .511, .286, and .203 respectively. The results suggest that the factors might be a part of a larger alienation construct in black delinquents.

TERMS such as hostility, aggressiveness, and sociopathic are often used to label black delinquents. To date, little empirical evidence exists which describes meaningful dimensions related to black delinquency. In a study of black adolescents, Shuman and Hatchett (1974) concluded that one of the most pervasive factors in black adolescents was alienation.

The absence of data and the results of Shuman and Hatchett's (1974) study appeared sufficiently promising to encourage further investigation. The purpose of the study was to identify meaningful personality dimensions among black delinquents within the Jesness Inventory (JI) (Jesness, 1966) as a first step to establishing the construct validity of the instrument.

*Method*

The subjects were 250 adjudicated, black delinquent males randomly selected from the North Carolina Youth Development Schools. The mean age was 14.6 years. All subjects were administered the JI in the youth development schools.

The JI is a 155 item, true-false inventory purporting to measure 10 personality characteristics of delinquents and deviant adolescents: (1) Social Maladjustment, (2) Value Orientation, (3) Immaturity, (4) Autism, (5) Alienation, (6) Manifest Aggression, (7) Withdrawal, (8) Social Anxiety, (9) Repression, and (10) Denial. Raw scores for each variable are converted to T scores.

A principal components factor analysis was computed from the matrix of product moment correlations. For the analysis, the total variance was factor analyzed by placing unities in the diagonals. Factors with eigenvalues greater than unity were extracted and rotated through a varimax solution.

*Results and Discussion*

The rotation of factors resulted in the extraction of three factors having corresponding eigenvalues of 4.09, 2.29, and 1.62. Major results of the factor analysis are presented in Table 1.

The first factor was termed Self-Estrangement and the proportion of the total variance extracted was .511. Scales heavily loaded on Factor I seem to describe the black delinquent as possessing antisocial tendencies, alienated feelings toward others, distorted perceptions of reality, a dissociation between events and the person, and feelings of frustration and anger. These behavioral characteristics in black delinquents

TABLE I  
*The Rotated Factor Matrix of the Jesness Inventory Scales*

Scales	Factor Loadings <sup>a</sup>			
	I	II	III	$h^2$
1. Social Maladjustment	.69	.61	—	.86
2. Value Orientation	.88	—	—	.90
3. Immaturity	—	—	.82	.78
4. Autism	.85	—	—	.77
5. Alienation	.86	—	—	.75
6. Manifest Aggression	.69	.48	—	.73
7. Withdrawal	—	.77	—	.65
8. Social Anxiety	—	.46	—	.35
9. Repression	—	—	.86	.77
10. Denial	.81	—	—	.68

<sup>a</sup> Loadings less than .40 omitted.

seem to be basic to attitudes of self-estrangement. The factor of Self-Estrangement may refer to black delinquents' beliefs that they are not what they would like to be.

The second factor was termed Social Isolation, and the proportion of total variance extracted was .286. Inspection of the scales on Factor II reveals that this dimension describes the black delinquent as withdrawn and unhappy, angry, and disturbed over interpersonal relationships. This factor may reflect black delinquents' feelings of isolation from family and society.

The third factor was termed Immaturity, and the proportion of total variance extracted was .203. Although only two scales had high loadings (Table 1) on Factor III, the dimension described black delinquents as repressing feelings and beliefs normally experienced by people and as displaying attitudes about themselves and others normally experienced by younger persons.

Results suggest that the personality dimensions associated with black delinquents are an indication of social and emotional maladjustment. The factors of Self-Estrangement, Social Isolation, and Immaturity may be dimensions of a larger more pervasive construct of alienation in black delinquents. The implication of the current study is that new research identify meaningful personality dimensions of black delinquents as well as those of white and Indian groups.

## REFERENCES

- Jesness, C. *The Jesness Inventory*. Palo Alto, California: Consulting Psychologists Press, 1966
- Shuman, H. and Hatchett, S. *Black racial attitudes: Trends and complexities*. Ann Arbor, Michigan: University of Michigan, 1974





## THE DEVELOPMENT AND VALIDATION OF THE STUDENT OPINION INVENTORY FACTOR SCALES<sup>1</sup>

JAN PERNEY  
Boston College

A description is given of the development of the Student Opinion Inventory (SOI), an instrument designed to measure attitudes of students in secondary schools toward several aspects of their schools. The purposes of this investigation were to determine the concurrent validity of the SOI factor scales and to examine the reproducibility of the reliability estimates found in pilot studies of the instrument. Responses to the SOI from 367 students indicated that 5 of the 6 factor scales of the SOI possessed some concurrent validity. Furthermore, the obtained reliabilities of the factor scales were relatively high for attitudinal measures and closely reproduced reliability estimates established by the final pilot study of the instrument.

ONE aspect which was seldom included in school evaluations until recently is the assessment of attitudes of students toward their schools. One reason that students' attitudes were not widely assessed may have resulted from the absence of an appropriate instrument to measure students' attitudes toward the entire school program. Attitudinal instruments developed prior to 1973 either gathered information on only one or two dimensions (Finch, 1969; Educational Testing Service, 1972) or furnished data too global to be of diagnostic value (Remmers, 1952). To provide an instrument which measured attitudes of students toward several aspects of their schools, the Student Opinion Inventory (National Study of School Evaluation, 1974) was developed under the auspices of the National Study of School Evaluation (NSSE). The purposes of this current investigation were to determine the concurrent validity of the Student Opinion Inventory (SOI) factor scales

<sup>1</sup> The author wishes to thank his doctoral committee and Dr. Peter Arasian for commenting on previous drafts of this paper.

and to examine the reproducibility of reliability estimates found in pilot studies of the instrument.

### *The Instrument*

The pilot SOI contained 48 5-point scale items written to correspond to sections in *Evaluative Criteria* (National Study of School Evaluation, 1970), a widely circulated publication of the NSSE used for school evaluation and accreditation purposes. Each of the six groups of items which corresponded to a section in *Evaluative Criteria* was hypothesized to constitute a factor scale for the final instrument. The groups of items were written to assess attitudes of students toward (a) their teachers, (b) their counselors, (c) the school administration, (d) the curriculum and instruction of their school, (e) cocurricular activities, and (f) school characteristics in general.

To date one local and one national pilot study of the SOI have been conducted. For the local pilot study the SOI was administered to students in three urban midwest high schools in April, 1972. A principal components factor analysis with orthogonal rotation of the factors was used to analyze 1164 student responses. Since there were 9 eigenvalues greater than 1.0, 9 factors were rotated; however, only 6 factors were considered interpretable. The factor analysis indicated that the factors emerged essentially as predicted and that the coefficient alpha reliability estimates of the factor scales ranged from .66 to .86, with a median reliability of .80. Thus, the factor scales were judged to possess adequate reliabilities to warrant further investigation. The pilot SOI was revised on the basis of the local pilot study data.

For the second pilot study, a random sample was drawn from all accredited secondary schools in the continental United States. The SOI was administered to 1157 students from 43 schools in the sample during February, 1973. The same method of analysis that was used in the local pilot study was employed to examine the responses in the national pilot study. The results indicated a factor structure similar to that found in the pilot test analysis. Furthermore, the estimated reliabilities of the factor scales remained substantial. Following are the 6 factors which were obtained from the second pilot study. A description of the factor, the number of items ( $n$ ) per factor, and the reliability estimate ( $r$ ) for each factor are given:

*Student-Teacher (ST)*. Perceptions held by students about their teachers' helpfulness in learning subjects;  $n = 7$ ;  $r = .82$ .

*Student-Counselor (SC)*. Students' expressed feelings concerning counselors' helpfulness in vocational, academic and personal matters;  $n = 5$ ;  $r = .81$ .

*Student-Administration (SA)*. Students' attitudes toward the treatment of students by the administration;  $n = 6$ ;  $r = .76$ .

*Student-Curriculum and Instruction (SI)*. Students' perceptions about the adequacy of the curriculum and quality of teaching;  $n = 5$ ;  $r = .75$ .

*Student-Participation (SP)*. Students' attitudes toward cocurricular activities and participation in school life;  $n = 5$ ;  $r = .69$ .

*Student-School Image (SS)*. Satisfaction expressed by students with school in general and pride in their school;  $n = 6$ ;  $r = .78$ .

The final form of the SOI consisted of 34 items determined by the factor analysis of the responses from the national pilot study. More information concerning the background and use of the SOI is contained in the manual for the instrument.

### *Methodology*

To investigate the concurrent validity of the first three SOI factor measures, semantic differential scales were constructed to reflect attitudes similar to those assessed by the ST, SC, and SA factor scales. For example, a semantic differential scale was constructed to measure how much students liked their counselors. The correlation between the semantic differential scale designed to portray student liking of counselor and the SC factor scale was expected to be moderate and positive. To investigate the validity for the two additional SI and SP factor scales, semantic differential scales which were intended to measure the same attitudes as those thought to be portrayed in the SI and SP factor scales were constructed. Since it was hypothesized that the same attitudes were being measured by both the SOI factor scales and semantic differential scales, it was expected that the validity coefficients would be positive and somewhat larger for the SI and SP scales than those for the ST, SC, and SA factor scales. Because of time constraints, no validity data were gathered for the SS factor scale.

### *Results and Discussion*

The results of the current study were based on responses of 367 students from seven secondary schools located throughout the United States. The coefficient alpha reliability estimates for the semantic differential scales ranged from .70 to .87 with a median reliability of .84. The correlations of the ST, SC, and SA factor scales with their semantic differential counterparts were .38, .49 and .36, respectively, whereas the reliabilities were .80, .81, and .75, respectively. Because in this case the SOI factor scales and the semantic differential scales were hypoth-

esized to be measuring similar, but not the same attitudes, the moderate positive correlations were expected. The validity coefficients for the SI and SP factor scales were .59 and .50 respectively, whereas the reliability estimates were .75 and .66, respectively. As expected, the correlations were positive and somewhat larger than were the correlations for the ST, SC and SA factor scales. Moreover, the reliability estimates for the factor scales in this current study closely approximated the reliability estimates determined in the national pilot administration of the instrument. For example, the obtained reliability for the SA factor scale in this current study was .75, whereas for the national pilot study the obtained reliability was .76.

### Conclusions

The following conclusions were formulated:

1. The factor scales of the SOI appeared to be sufficiently reliable. Not only were the internal consistency reliability estimates relatively high for attitudinal measures, but also the reliability estimates found in this current study closely approximated those obtained in the national pilot study.
2. The factor analysis of the responses in the pilot tests provided validity evidence in two ways. First, the factors which emerged in the pilot test were predicted when the instrument was constructed. Second, similar factor structures were obtained in both the local and the national pilot tests.
3. Further validity evidence was established by this current study, since the correlations predicted to exist between the SOI factor scales and the semantic differential scales were obtained.

### REFERENCES

- Educational Testing Service. *Student Instructional Report*. Princeton, New Jersey: Educational Testing Service, 1972.
- Finch, C. R. Instrument to assess student attitude toward instruction. *Journal of Educational Measurement*, 1969, 6, 257-258.
- National Study of School Evaluation. *Evaluative Criteria* (4th ed.). Arlington, Va.: National Study of School Evaluation, 1970.
- National Study of School Evaluation. *Student Opinion Inventory: Manual*. Arlington, Va.: National Study of School Evaluation, 1974.
- Remmers, R. R. Attitude toward school. In M. Shaw and J. Wright (Eds.), *Scales for the Measurement of Attitudes*. New York: McGraw-Hill, 1967.

# THE RELATIONSHIP OF READING ACHIEVEMENT, SCHOOL ATTITUDE, AND SELF-RESPONSIBILITY BEHAVIORS OF SIXTH-GRADE PUPILS TO COMPARATIVE AND INDIVIDUALIZED REPORTING SYSTEMS: IMPLICATIONS FOR IMPROVEMENT OF VALIDITY OF THE EVALUATION OF PUPIL PROGRESS

THOMAS W. BUTTERWORTH

Los Angeles County Schools

WILLIAM B. MICHAEL

University of Southern California

Two comparable samples of sixth-grade pupils were recipients of information furnished by two different systems of reporting pupil progress: (a) one involving use of a traditional competitive and comparative A-F letter grade approach and (b) the other embracing a highly individualized procedure which consisted of detailed narrative statements providing evaluative feedback on performance in the school setting. A  $2 \times 2 \times 2$  quasi-experimental design (reporting system  $\times$  IQ  $\times$  sex) was employed with dependent variables including measures of (a) reading achievement, (b) school attitude, and (c) self-responsibility for intellectual attainments. Results from three univariate analyses of variance revealed significant main effects for each dependent variable favoring the individualized reporting system over the traditional one, high ability children over low ability children, and girls over boys. In the instance of the measure of intellectual self-responsibility a significant interaction occurred between ability level and mode of reporting which suggested that as compared with the traditional competitive mode an individualized reporting system would yield differential outcomes indicating a higher average level of intellectual self-responsibility for children of low ability but no appreciable difference in average level of self-responsibility for children of high ability. Implications of the use of individualized reporting systems for improving the validity of evaluating pupil progress are discussed.



FOR two samples of 300 sixth-grade pupils (600 children altogether) from two different adjacent Southern California suburban communities of a comparable socio-economic level and similar ethnic composition the purpose of this investigation was to determine whether differences in reading achievement, attitude toward school, and self-responsibility (dependent variables) were related to (a) employment of either one of two systems of reporting pupil progress—one constituting a traditional comparative A to F letter grade approach and the other an individualized procedure involving an evaluative feedback in the form of narrative statements without symbols (treatment factor), (b) high or low general ability (as determined respectively by placement at an IQ of 100 or above or falling at an IQ below 100 in terms of a deviation IQ based on the total raw score on the five-test Verbal Scale of the Lorge-Thorndike Intelligence Tests, 1964, Multi-Level Edition, Level D, Form 1), and (c) sex. Interest also centered on possible interactions between the treatment factor and either general ability level or sex with respect to either reading performance or affective dimensions of school attitude or self-responsibility. Because of the paucity of published research, the data from this study could possibly provide guidelines for developing report systems that would have augmented validity in the evaluation of pupil progress.

### *Methodology*

In the data analysis a quasi-experimental  $2 \times 2 \times 2$  factorial design (treatment  $\times$  IQ  $\times$  sex) (Kirk, 1968, pp. 221-227) was used to examine main effects and interaction effects associated with posttreatment scores on each of three dependent variables: (a) reading achievement as indicated by standing on the Comprehensive Test of Basic Skills (CTBS), Reading (Level 2, Form R), total score, (b) attitude toward school as revealed by responses to a Semantic Differential (SD) scale consisting of 12 school concepts each permuted with five bipolar adjectives (evaluative in emphasis), and (c) self-responsibility as defined by placement on the Intellectual Achievement Responsibility (IAR) measure (Crandall, Katkovsky, and Crandall, 1965).

Pretreatment equivalence of the two samples studied was demonstrated by a lack of significant differences in pretreatment mean scores on each of the three dependent variables as well as in mean IQ scores which to the nearest integer were 101 and 102 for the two samples.

### *Findings*

The major results of posttreatment testing, which are set forth in Tables 1 and 2, may be summarized as follows:

TABLE 1

*Summary of Analyses of Variance for Reading Achievement Scores on the CTBS, Attitude Scores on the SD, and Self-Responsibility Scores on the IAR*

Source of Variation	df	Reading Achievement (CTBS)		Attitude toward School (SD Measure)		Self-Responsibility (IAR Measure)	
		MS	F	MS	F	MS	F
Reporting System							
(RS)	1	15.55	7.44**	50912.85	21.92**	301.04	16.53**
Sex (S)	1	8.21	3.93*	50178.59	21.60**	217.20	11.93**
Ability (A)	1	1134.91	543.02**	20638.90	8.88**	312.48	17.16**
RS × S	1	0.04	—	539.60	—	3.08	—
RS × A	1	1.23	—	2683.99	1.16	126.04	6.92**
S × A	1	0.49	—	1016.66	—	2.54	—
RS × S × A	1	0.02	—	2472.56	1.06	1.98	—
Within	592	2.09	—	2323.11	—	18.21	—

Subgroup	CTBS Means	SD Means	IAR Means*
Comparative Reporting	6.82	173.24*	24.71
Individualized Reporting	7.14	154.82	26.13
Boys	6.86	173.17	24.82
Girls	7.10	154.82	26.02
High Ability	8.36	158.16	26.14
Low Ability	5.61	169.89	24.70

\*  $p < .05$ .\*\*  $p < .01$ .

\* Lower number indicative of more positive attitude.

1. On each of the three dependent variables statistically significant main effects occurred relative to mode of reporting system employed, ability level, and sex. In particular, the mean scores on the reading and self-responsibility measures were observed to be higher for the individual reporting system than for the competitive A to F marking system, for boys than for girls, and for high ability than for low ability children. On the SD measure of attitude toward school the mean scores were lower (indicating a more favorable attitude) for the subgroup exposed to the individual reporting sys-

TABLE 2

*Mean Self-Responsibility Scores of the IAR in the Statistically Significant Reporting System by Ability Interaction*

Reporting System	Ability		Difference between $A_1$ and $A_2$ Means (Column 1 minus Column 2)
	High ( $A_1$ )	Low ( $A_2$ )	
Comparative ( $RS_1$ )	25.89	23.53	2.36
Individualized ( $RS_2$ )	26.39	25.87	0.52
Differences between $RS_1$ and $RS_2$ means (Row 1 minus Row 2)	0.50	2.34	

- tem than for the subgroup exposed to the competitive reporting system, for girls than for boys, and for the high ability level than for the low ability level subgroups. In terms of two-tailed  $t$ -tests all differences were significant beyond .01 level except for the single difference between girls and boys on the reading achievement variable which was significant at the .05 level (but still at the .01 level for the directional null hypothesis that the mean of the population of girls would be less than or equal to that of the boys).
2. The only statistically significant interaction effect ( $p < .01$ ) was that between treatment (type of reporting system) and intellectual ability level for the dependent variable concerned with self-responsibility. An inspection of the entries in Table 2 reveals that in the instance of the comparative reporting system ( $RS_1$ ) a difference of 2.36 occurred between the means of 25.89 and 23.53 for the high ability ( $A_1$ ) and low ability ( $A_2$ ) subgroups, respectively, or that in the case of the low ability ( $A_2$ ) subgroup a difference of 2.34 occurred between the means of 25.87 and 23.53 for those subgroups exposed, respectively, to the individualized ( $RS_2$ ) and comparative ( $RS_1$ ) reporting systems. (Although not reported,  $F$  ratios—associated with one and 592 degrees of freedom—of 22.95 for simple effects between  $A_1$  and  $A_2$  at the  $RS_1$  level and of 22.43 for simple effects between  $RS_2$  and  $RS_1$  at the  $A_2$  level were significant considerably beyond the .01 level.)

### Conclusions

The following conclusions may be formulated:

1. It would appear that the individualized reporting system as compared with the competitive A to F reporting system was associated with higher reading performance, a more favorable attitude toward school, and a greater sense of self-responsibility.
2. Although pupils of relatively high ability differed little in their level of self-responsibility irrespective of how their progress in school was evaluated and reported, children of relatively low ability displayed a higher level of self-responsibility when their work was judged and reported in an individualized manner with narrative statements embodying feedback than when their work was evaluated and reported in a traditional competitive A to F symbolic format. Apparently attitude toward school as measured as well as cognitive performance in reading were not dependent upon interaction effects between ability level and mode of reporting pupil progress.

*Discussion: Implications for Improving Validity of  
the Evaluation of Pupil Progress*

It would appear that the feedback provided by detailed narrative statements in reports of pupil attainments could have accounted in large part for the higher average standings in the reading measure. In turn it would not seem unexpected that as cognitive performances improve, both more favorable attitudes toward school and a heightened sense of responsibility for one's own accomplishments would occur. In the instance of the interaction between ability level and mode of reporting pupil performance it appeared that pupils of lower ability levels as compared with those at higher ability levels might have been aided to a greater degree in their acquisition of a heightened level of self-responsibility. With greater lapse in time similar interactive effects might have taken place relative to reading achievement and attitude-toward-school variables. From the standpoint of improving the validity of evaluation of pupil progress the individualized report system as compared with the traditional comparative and competitive A to F system would appear to have the following advantages: (a) specificity of feedback regarding mastery of designated course objectives, (b) means for on-going formative evaluation permitting changes in curriculum and in instructional strategies in view of improved diagnosis of areas of strength and weakness, and (c) improved basis for communicating to parents and pupils those cognitive and affective behaviors that might be modified to facilitate pupil growth and self-actualization. Such facilitative effects associated with individualized reporting systems may be relatively more important for children of lower levels of ability than for those of higher levels of ability.

#### REFERENCES

- Crandall, V. C., Katkovsky, W., and Crandall, V. J. Children's beliefs in their own control of reinforcements in intellectual-academic achievement situations. *Child Development*, 1965, 36, 91-109.
- Kirk, R. E. *Experimental design: Procedures for the behavioral sciences*. Belmont, Calif. Brooks/Cole, 1968.





INTERRELATIONSHIPS AMONG 76 INDIVIDUALLY-  
ADMINISTERED TESTS INTENDED TO REPRESENT 76  
DIFFERENT STRUCTURE-OF-INTELLECT ABILITIES AND A  
STANDARDIZED GENERAL INTELLIGENCE TEST IN A  
SAMPLE OF 34 NINE-YEAR-OLD CHILDREN<sup>1,2</sup>

JANE FAVERO AND JULE DOMBROWER

Glendora Unified School District

WILLIAM B. MICHAEL AND LEO RICHARDS

University of Southern California

For a sample of 34 nine-year-old children from a southern California middle-class suburban community both the scores of 76 *individually* administered structure-of-intellect (SOI) tests constructed or selected to duplicate exactly 76 hypothesized SOI abilities and the scores on the verbal (V), nonverbal (NV), and composite (C) scales of the Lorge-Thorndike Intelligence Tests (LT), Multi-Level Edition, were intercorrelated and factor analyzed to determine the extent of overlap of SOI ability measures, their degree of relationship with the LT scales, and the possible presence of second order factors among the SOI tests. Although the data revealed a range in magnitude of the 2850 correlation coefficients among the SOI measures from  $-.47$  to  $.69$  (median coefficient,  $.13$ ), the values for the ranges and the average magnitudes of intercorrelation coefficients of SOI tests within *single categories of the same* operations, contents, or products dimension of the SOI model did not differ appreciably from those corresponding values and magnitudes found *between categories from dif-*

<sup>1</sup> This report is based on research findings from a Title VI B project, Project Number 19 64576-4123-02, entitled "Learning How to Learn," funded by the Department of Health, Education, and Welfare, United States Office of Education.

<sup>2</sup> For their invaluable assistance and contributions appreciation is expressed to the following teachers: Wanda Bell, Corrine Daum, Janet Dean, Raymond Fuller, Phyllis Huston, Edie Johnson, Patricia Keith, Patricia Petridis, Elsie Westlake, and Helen Lorenz; to the following aides: Laurel Carter, Lois Christensen, Janice Hagey, Rosemary Henley, Janet McNary, Irene Mieger, Jean Sturm, Jacqueline Toigo, Patricia Wright, and Margaret King; to Nancy Weingartner, school psychologist; to Mary Ann Poole, clerk typist; and to Maxine Pennington, manuscript typist.

*ferent* SOI dimensions. In the absence of identifiable meaningful second order factors or dimensions, there was, however, the suggestion of a factor of general intellectual function in view of the high loadings of several SOI tests on the same factor as that on which the LT-V and LT-NV scales were heavily saturated. A weighted combination of eight to ten SOI ability tests could afford a potentially valid representation of the complex of functions or of the general function being measured by the LT-V and LT-NV scales.

For a sample of 34 children who resided in a middle-class suburban community in southern California the purposes of this descriptive correlational investigation were (1) to ascertain the degree of interrelationship among 76 individually-administered tests designed to duplicate exactly 76 factors in the Structure-of-Intellect (SOI) model (Guilford, 1967), (2) to determine the extent of relationship of each of these SOI tests with the verbal, nonverbal, and composite scales of the 1964 Lorge-Thorndike Intelligence Tests, Multi-Level Edition (LT-V, LT-NV, and LT-C), and (3) to obtain evidence regarding the nature of second-order factors, if any, that would describe the interrelationships among the measurable constructs in the SOI model. The research to be reported is an outgrowth of a school district project that was initiated to provide individualized instruction to children in developing increased competencies in their use of several SOI abilities. It was thought that the correlational data would furnish important information concerning the degree of overlap among SOI tests as well as an indication of the extent to which such tests are related to an intended measure of so-called general intelligence. In addition, evidence could also be obtained to determine whether certain categories within each of the three broad dimensions of operations, contents, and products in the SOI model tend to be highly interrelated and thus possibly indicative of higher order constructs of intellectual function.

### *Methodology*

#### *Sample*

The sample of 34 children of whom there were 12 boys and 22 girls, ranged in age from 9 years 0 months to 9 years 11 months. Only those pupils were selected whose deviation IQ scores on the composite scale of the LT fell between 85 and 115.

#### *Tests*

To furnish a measure of general intellectual status the LT Intelligence Tests, Multi-Level Edition, Levels A and B, Form 1 were

administered approximately 3 to 12 months prior to the time that the children were given the SOI tests. Deviation IQ scores were determined on the previously cited LT-V, LT-NV, and LT-C scales.

Through use of models of test items provided by Guilford (1967), Guilford and Hoepfner (1971), Meeker (1969), and reports from the Aptitudes Research Project at the University of Southern California, a committee of teachers, teacher aides, and school psychologists, who had had special training in a workshop during the previous summer, constructed and in a few instances selected tests to represent the 76 SOI abilities. These tests were carefully reviewed and edited and in many instances subsequently revised by psychologists associated with the school project.

To portray the five operations (inferred psychological processes) of cognition (*C*), memory (*M*), evaluation (*E*), convergent production (*N*), and divergent production (*D*), 16, 14, 16, 15, and 15 tests, respectively, were devised or selected. In the contents dimension, the numbers of these same tests to represent the figural (*F*), symbolic (*S*), and semantic (*M*) categories of given information to be processed were, respectively, 25, 23, and 28. For the six products (new information arising from processing of given information) of units (*U*), classes (*C*), relations (*R*), systems (*S*), transformations (*T*), and implications (*I*), the corresponding frequencies of tests employed were 13, 15, 14, 10, 15, and 9.

In Table 1 each of the 76 SOI tests is designated in terms of the trigram notation employed by Guilford (1967) to describe an hypothesized ability factor. The first letter refers to the operations category; the second, to the contents category; and the third, to the products category. One can interpret a trigram as revealing what *operation* is being used to process given information (*content*) to bring about new information (*product*)—that is, *operating on content* to obtain a *product*. For example, the trigram CFS, which is cognition of figural systems, indicates that an examinee would use the operation of cognition to process figural content to obtain a product in the form of a system.

Because of the exploratory nature of this investigation and because of the somewhat limited size of the sample, it did not appear feasible to undertake the almost prohibitive amount of expense and time required to obtain estimates of the reliability of the 76 SOI tests. In a few instances some commercially available tests were employed to duplicate a given SOI factor. An effort was made to approximate tests that had been used in the Aptitudes Research Project of the University of Southern California, but to adapt the difficulty level. Thus the estimates of reliability provided by Guilford and Hoepfner (1971) would

TABLE I  
*Trigram Designation of and Number of Items in Each of the 76 SOI Test Variables Along with the Correlations of Each SOI Test with the Large-Thorndike Intelligence Tests, Multi-Level Edition—Verbal Scale (LT-V), Nonverbal Scale (LT-NV), and Composite Scale (LT-C) (Decimal Points Omitted)*

SOI Test Variable Number	SOI Test Variables Trigram Designation	Max. Score Correlation with			SOI Test Variable Number	SOI Test Variables Trigram Designation	Max. Score Correlations with		
		LT-V (1)	LT-NV (2)	LT-C (3)			LT-V (1)	LT-NV (2)	LT-C (3)
4	CFU	20	-14	-14	42	ESS	12	42	42
5	CFC	15	49	51	43	EST	12	45	43
6	CFR	14	28	30	44	EMU	15	10	11
7	CFS	26	06	11	45	EMC	15	53	53
8	CFT	13	42	55	46	EMR	13	43	43
9	CFI	14	26	32	47	EMS	08	00	04
10	CSU	24	33	35	48	EMT	12	-05	-06
11	CSC	16	51	59	49	EMI	16	11	11
12	CSR	12	45	48	50	NFC	12	33	37
13	CSS	15	45	51	51	NFR	15	45	48
14	CST	12	43	49	52	NFT	27	10	13
15	CMC	19	22	21	53	NFI	09	17	17
16	CMR	15	22	23	54	NSU	16	31	33
17	CMS	15	34	37	55	NSC	14	23	19
18	CMT	59	08	10	56	NSR	14	14	13

19	CMI	27	47	42	45	57	NSS	14	60	62	62
20	MFU	16	26	28	27	58	NST	17	23	29	26
21	MFC	15	16	18	18	59	NMU	12	20	22	23
22	MFR	15	54	62	59	60	NMC	12	-15	-16	-17
23	MFT	15	24	28	26	61	NMR	12	36	37	37
24	MEI	15	01	03	02	62	NMT	17	39	32	31
25	MSU	16	24	27	24	63	NMI	12	-04	06	01
26	MSC	15	13	25	19	64	DFU	OE <sup>a</sup>	20	05	13
27	MSR	15	26	26	27	65	DFC	OE <sup>a</sup>	50	52	50
28	MST	15	25	18	21	66	DFS	OE <sup>a</sup>	08	07	09
29	MMU	20	12	14	15	67	DFT	OE <sup>a</sup>	43	42	43
30	MMC	14	-08	-09	-08	68	DFI	OE <sup>a</sup>	-03	-05	-06
31	MMR	15	10	09	09	69	DSU	OE <sup>a</sup>	23	16	20
32	MMT	21	31	34	34	70	DSC	OE <sup>a</sup>	30	26	28
33	MMI	17	-01	02	02	71	DSR	OE <sup>a</sup>	37	40	39
34	EFU	14	08	15	12	72	DST	OE <sup>a</sup>	50	40	45
35	EFC	12	16	26	20	73	DMU	OE <sup>a</sup>	-28	-29	-30
36	EFR	12	69	74	72	74	DMC	OE <sup>a</sup>	17	15	16
37	EFS	12	35	38	37	75	DMR	OE <sup>a</sup>	38	37	39
38	EFT	11	-05	-13	-08	76	DMS	OE <sup>a</sup>	-12	-07	-10
39	ESU	20	14	13	13	77	DMT	OE <sup>a</sup>	22	12	18
40	ESC	12	39	40	41	78	DMI	OE <sup>a</sup>	-15	-18	-16
41	ESR	20	65	54	60	79	NMS	10	12	18	17

<sup>a</sup> OE—open-ended test with no maximum number of items or scores.



not be entirely appropriate. Communality estimates arising from the factor analysis of the correlational data as well as the highest correlation which any test exhibited with another test in the correlation matrix could probably be taken as lower bound approximations to reliability. In general, it appeared that virtually every SOI test would yield an estimate of reliability in excess of .40 and usually an estimate greater than .50.

### *Statistical Analysis*

Pearson product-moment coefficients of correlation were found between sets of scores on all SOI tests and on the LT-V, LT-NV, and LT-C scales. These correlational data are summarized descriptively in Tables 1 through 5. The only inferential interpretation that could be rendered would be in relation to the correlation coefficients presented in Table 1. For a sample of 34 subjects, coefficients of .34 and .44 are required for significance at the .05 and .01 levels, respectively. Significance tests of median values of a distribution of correlated or dependent correlation coefficients were not known to the writers. Hence, the correlational data in Tables 2, 3, 4, and 5 could be treated only descriptively.

A principal components factor analysis followed by rotation to satisfy the Varimax criterion was completed. Because of the presence of more test variables than of subjects, unities were inserted in the diagonals of the correlation matrix, although an analysis based upon the insertion of the highest column correlation coefficient for the diagonal entry would not be expected to alter the basic factor structure to a substantial degree. Only those rotated factors which had at least three variables with loadings equal to or greater than .35 were examined for possible interpretation.

### *Findings*

In relation to the first objective of the investigation concerning the intercorrelations among the 76 SOI tests, the following outcomes may be summarized:

1. The range in values of the 2850 correlation coefficients was from  $-.47$  to  $.69$ , and the median coefficient was equal to  $.13$ .
2. For the dimension of operations, the data for which are summarized in Table 2, the median coefficients within the five categories and between the 10 pairings of categories were, respectively,  $.18$  and  $.11$ . Within categories, the highest median coefficient of  $.28$  was for tests of cognition, and the lowest median coefficient of

TABLE 2

Permutations of Five SOI Operations Categories: Number of Tests in Each Category, Number of Test Intercorrelations within Each Category (Diagonal Entries) or between Each Pairing of Categories (Off Diagonal Entries), Median Values of These Intercorrelations, and Their Ranges<sup>a</sup>

Categories and Numbers of Tests	(1) Cognition (C) 16 Tests	(2) Memory (M) 14 Tests	(3) Evaluation (E) 16 Tests	(4) Convergent Production (N) 15 Tests	(5) Divergent Production (D) 15 Tests
(1) Cognition (C) 16 Tests	(120) .28 .24 to .67	(224) .17 -.32 to .57	(256) .20 -.30 to .63	(240) .25 -.35 to .68	(240) .11 -.44 to .65
(2) Memory (M) 14 Tests	(224) .17 -.32 to .57	(91) .11 -.28 to .47	(224) .08 -.46 to .52	(210) .14 -.46 to .60	(210) .01 -.39 to .48
(3) Evaluation (E) 16 Tests	(256) .20 -.30 to .63	(224) .08 -.46 to .52	(126) .18 -.26 to .52	(240) .17 -.29 to .60	(240) .11 -.38 to .56
(4) Convergent Production (N) 15 Tests	(240) .25 -.35 to .68	(210) .14 -.46 to .60	(240) .17 -.29 to .60	(105) .18 -.26 to .69	(225) .08+ -.41 to .51
(5) Divergent Production (D) 15 Tests	(240) .11 -.44 to .65	(210) .01 -.39 to .48	(240) .11 -.38 to .56	(225) .08 -.47 to .51	(105) .16+ -.09 to .58

<sup>a</sup> In this table and in the two tables to follow, the first row of each cell indicates the number of intercorrelations of tests (entries in parentheses); the second row, the median value of the intercorrelations; and the third row, the range in values of the intercorrelations.

.11 was for tests of memory. Between pairings of categories the two highest median coefficients of .25 and .20 were for the respective permutations of cognition and convergent production and of cognition and evaluation; the lowest median coefficient of .01 was associated with the pairing of divergent production with memory.

- For the dimension of contents, the data for which are set forth in Table 3, the median coefficients within the three categories and between the three pairings of categories were, respectively, .15 and .15. Within categories, the greatest median coefficient of .20 was for tests involving symbolic content, and the lowest median coefficient of .13 was for measures of figural stimuli. Between pairings of categories, the highest median coefficient of .16 was for the permutation of figural material with symbolic content; the lowest median coefficient of .12 arose in conjunction with the permutation of figural and semantic items.
- For the dimension of products, the data for which are presented in Table 4, the median coefficients within the six categories and between the 15 pairings of categories were, respectively, .12 and .14. Within categories, the highest median coefficient of .28 was

TABLE 3

*Permutations of Three of the Four SOI Contents Categories (Behavioral Omitted), Number of Tests in Each Category, Number of Test Intercorrelations within Each Category (Diagonal Entries) or between Each Pairing of Categories (Off Diagonal Entries), Median Values of These Intercorrelations and Their Ranges*

Categories and Numbers of Tests	(1) Figural (F) 25 Tests	(2) Symbolic (S) 23 Tests	(3) Semantic (M) 28 Tests
(1) Figural (F) 25 Tests	(300) .13 -.31 to .58	(575) .16 -.47 to .69	(700) .12 -.46 to .60
(2) Symbolic (S) 23 Tests	(575) .16 -.47 to .69	(253) .20 -.26 to .64	(644) .15 -.42 to .68
(3) Semantic (M) 28 Tests	(700) .12 -.46 to .60	(644) .15 -.42 to .68	(378) .15 -.46 to .60

TABLE 4

*Permutations of Six SOI Products Categories: Number of Tests in Each Category, Number of Test Intercorrelations within Each Category (Diagonal Entries) or between Each Pairing of Categories (Off Diagonal Entries), Median Values of Their Intercorrelations and Their Ranges*

Categories and Number of Tests	(1) Units (U) 13 Tests	(2) Classes (C) 15 Tests	(3) Relations (R) 14 Tests	(4) Systems (S) 10 Tests	(5) Transformations (T) 15 Tests	(6) Implications (I) 9 Tests
(1) Units (U) 13 Tests	(78) .11 -.28 to .47	(195) .09 -.31 to .56	(182) .10 -.32 to .52	(130) .13 -.35 to .58	(195) .07 -.47 to .54	(117) .08 -.46 to .58
(2) Classes (C) 15 Tests	(195) .09 -.31 to .50	(105) .11 -.28 to .55	(210) .22 -.35 to .65	(150) .14 -.33 to .64	(225) .14 -.39 to .56	(135) .07 -.31 to .64
(3) Relations (R) 14 Tests	(182) .10 -.32 to .52	(210) .22 -.35 to .65	(91) .28 -.04 to .56	(140) .19 -.21 to .69	(210) .22 -.22 to .68	(126) .14 -.30 to .58
(4) Systems (S) 10 Tests	(130) .13 -.35 to .58	(150) .14 -.33 to .64	(140) .19 -.21 to .69	(45) .13 -.18 to .55	(150) .14 -.31 to .56	(90) .14 -.29 to .58
(5) Transformations (T) 15 Tests	(195) .07 -.47 to .54	(225) .14 -.39 to .56	(210) .22 -.22 to .68	(150) .14 -.31 to .56	(105) .12 -.39 to .38	(135) .11 -.44 to .58
(6) Implications (I) 9 Tests	(117) .08 -.46 to .58	(135) .07 -.31 to .49	(126) .14 -.30 to .50	(90) .14 -.29 to .52	(135) .11 -.44 to .49	(36) .08 -.32 to .58

for tests involving relations, and the lowest median coefficient of .08 was for tests representing implications. Between pairings of categories the three highest median coefficients of .22, .22, and .19 were for the respective pairings of classes with relations, relations with transformations, and relations with systems; the two lowest median coefficients (both .07) were associated with the pairings of units with transformations and of classes with implications.

Relative to the second objective of the study which was concerned with the degree of relationship of each of the SOI tests with the V, NV, and C scales of the LT, complete data are provided in Table 1, and summary information of the frequency distributions of the correlation coefficients between deviation IQ scores on the LT-C scale and scores on the SOI tests for each of the categories from the operations, contents, and products dimensions is set forth in Table 5. In view of the presence of a correlation of .94 between the LT-V and LT-NV scales, only data pertaining to the correlation of SOI tests with the LT-C scale are described in Table 5. The following results may be summarized:

1. From the entries in Table 1, it is apparent that the three SOI tests with the highest coefficients of correlation with the composite

TABLE 5

*Frequency Distributions of Correlation Coefficients between IQ Scores on the Composite Scale of the Large-Thorndike Intelligence Tests, Multi-Level Edition, and Scores on SOI Tests Designed to Represent Each of the Categories within the Operations, Contents, and Products Dimensions*

Class Intervals for Correlation Coefficients ( $r$ 's)	Categories within Dimensions														
	Operations					Contents					Products				
	C	M	E	N	D	F	S	M	U	C	R	S	T	I	
.70 to .79			1			1					1				
.60 to .69			1	1			2				1	1			
.50 to .59	4	1	1	0	1	4	2	1		4	1	1	1		
.40 to .49	3	0	4	1	2	2	6	2		1	3	1	4	1	
.30 to .39	4	1	1	4	2	4	3	5	2	1	4	2	2	1	
.20 to .29	2	5	1	2	2	3	6	3	4	3	2	0	3	0	
.10 to .19	2	3	4	5	3	7	4	7	5	4	1	2	3	2	
.00 to .09	0	2	1	1	1	2		4	0	0	1	2	0	3	
.00 to -.09	0	2	2	0	1	2		2	1	1		0	2	1	
-.10 to -.19	1			1	2			3	0	1		1		1	
-.20 to -.29					0			0	1						
-.30 to -.39					1			1							
Number of Tests	16	14	16	15	15	25	23	28	13	15	14	10	15	9	
Median $r$ Value	.36	.20	.29	.23	.18	.23	.35	.16	.15	.24	.39	.27	.26	.02	
Range in $r$ Values	-.14	-.09	-.08	-.17	-.30	-.08	-.13	-.30	-.30	-.17	.09	-.16	-.08	-.16	
	to	to	to	to	to	to	to	to	to	to	to	to	to	to	
Number of $r$ 's $\geq .50$	.54	.51	.72	.62	.50	.72	.62	.53	.35	.54	.72	.02	.50	.45	
	4	1	3	1	1	5	4	1	0	4	3	2	1	0	

scale of the LT were EFR (.72), NSS (.62), and ESR (.60). The SOI test with the lowest coefficient was DMU with a value of  $-.30$ .

2. From the information cited in Table 5 it is evident that the highest coefficients appeared between the LT-C scale and those SOI tests reflecting the two operations of cognition and evaluation, content designated as symbolic, and the products of relations, classes, systems, and transformations. Correspondingly, the lowest levels of association occurred for the SOI tests involving the operation of divergent production, semantic content, and products of implications and units.

For the third objective of this study which was directed toward identifying possible second order factors (dimensions) underlying the intercorrelations of the 76 SOI tests, no tabular data are presented in view of the limitations in space and in light of the absence of readily interpretable results. The following outcomes are summarized:

1. Although 22 rotated factors with three or more loadings equal to or exceeding .35 emerged, the psychological meaningfulness of these factors was difficult to establish, as no definitive, clear-cut cluster or pattern of SOI tests representing common hypothesized characteristics appeared without the presence of another cluster or pattern affording a contradictory or alternative interpretation.
2. Perhaps the only factor that could be interpreted, at least tentatively, was the first one which yielded loadings on 22 SOI tests greater than or equal to .35 and which exhibited weights on 8 SOI tests of .50 or higher as evidenced by these SOI tests and their corresponding factor saturations: NSS, .73; EFR, .68; MFR, .65; ESR, .61; NFR, .55; CFC, .54; EMC, .53; and CFT, .50. Thus it would appear that a general intellectual factor was present, which embraced mostly operations of convergent production, evaluation, and cognition upon primarily figural and symbolic content resulting in new information or products largely in the form of classes, relations, and systems. The general factor interpretation was further supported by the presence of loadings of .94 on both the LT-V and LT-NV scales.

### *Conclusions*

Although there were marked fluctuations in the intercorrelations among the 76 SOI tests which could be attributed in part to the small sample size and to the limited degree of reliability of several of the shorter SOI tests, the following conclusions seem to be justified:

1. Some indication of the stability in the measured representation of



- a given SOI category within any one of the three dimensions of operations, contents, or products was evident from the small but positive value for the median correlation coefficient between SOI tests within that category.
2. Furthermore, the size of the median coefficient between SOI tests from different categories within the same dimension suggested that positive though quite modest relationships existed between different categories in each of the three dimensions of operations, contents, and products.
  3. In general, the average magnitudes of the interrelationships of SOI tests from the same category within a given dimension did not differ appreciably from those of the interrelationships of SOI tests from different categories within the same dimension. Thus for any one of the three SOI dimensions, tests designed to measure the same characteristic in that dimension exhibited intercorrelations of about the same magnitude as those for tests devised to measure different characteristics in that same dimension.
  4. That several of the SOI tests intended to represent quite different abilities showed substantial correlations with the LT scales as well as large loadings on the same factor or dimension as that on which the two LT subscales were heavily weighted suggests either that the LT is a highly complex instrument factorially, or that a general factor of intellectual function or of test-taking strategies might indeed exist. Additional efforts involving variations in factor analytic procedures might furnish evidence of a meaningful structure of second order factors which this factor analysis failed to reveal. (Incidentally, it should be mentioned that for the five subtests of vocabulary, sentence completion, arithmetic reasoning, verbal classification, and verbal analogies in the LT-V scale and for the three subtests of figure analysis, number series, and figure classification in the LT-NV scale the writers hypothesized that the respective factors would be CMU, EMS or CMS (depending on the maturity of the child and the level of item difficulties), CMS, CMC, CMR or EMR (depending on the maturity of the child and the level of item difficulties), CFR, CSS, and CFC. These hypothesized factors did not correspond too closely to those for the eight SOI tests that yielded the previously enumerated loadings in excess of .50 on the first factor on which both the LT-V and LT-NV scales were weighted .94.

### *Recommendations*

From a developmental point of view this study needs to be replicated with a larger group of children than was possible in this in-

vestigation and at different age levels, as the pattern of inter-correlations and the factor structure of measures of SOI abilities are probably related to the maturity levels and experiential backgrounds of examinees. In particular, a combined longitudinal-experimental investigation is recommended in which one of two large comparable groups of children does receive formal training in the development of SOI abilities and the second group does not. Application of parallel instrumentation at different time points would afford an indication of how interrelationships among measures of SOI abilities change as a function both of chronological age and of the amount of formal exposure to learning experiences intended to enhance the acquisition of these abilities.

### REFERENCES

- Guilford, J. P. *The nature of human intelligence*. New York: McGraw-Hill, 1967.
- Guilford, J. P. and Hoepfner, R. *The analysis of intelligence*. New York: McGraw-Hill, 1971.
- Meeker, M. N. *The structure of intellect: Its interpretation and uses*. Columbus, Ohio: Charles E. Merrill, 1969.

## THE RELATIONSHIP OF ACHIEVEMENT ON A TEACHER-MADE MATHEMATICS TEST OF COMPUTATIONAL SKILLS TO TWO WAYS OF RECORDING ANSWERS AND TO TWO WORKSPACE ARRANGEMENTS

GENE W. MAJORS

Anaheim Union High School District

JOAN J. MICHAEL

California State University, Long Beach

The performance of one sample of 120 seventh-grade students and of one sample of 120 eighth-grade students on a 30-item teacher-made test of computational skills was related to (a) transcribing the item responses to numbered spaces in an answer column on a separate sheet vs. writing the item response on the test form itself and (b) providing workspace on the test form itself vs. not providing such space. From the use of a  $2 \times 2$  quasi-experimental design, statistically significant differences were obtained with respect to both main effects ( $p < .05$ ) for the sample of seventh-grade students and with respect to the single main effect of mode of recording item response ( $p < .001$ ) for the sample of eighth-grade students. When the data were interrupted descriptively rather than inferentially there was the strong suggestion that recording answers directly on the examination form was associated with a higher average level of student performance than that realized when a detached answer column was used. For the sample of seventh-grade students, provision of working space on the test form was observed to yield higher average scores than when such a provision was not made.

FOR one sample of 120 seventh-grade students in four separate mathematics classes and for a second sample of 120 eighth-grade students also in four separate mathematics classes in a junior high school which is located in a mixed socioeconomic community in Orange County, California, the purpose of the investigation was to deter-

mine whether average level of achievement in a 30-item teacher-made test of arithmetic computation was related to (a) the presence of a detached answer column appearing on a separate sheet of paper and containing numbered spaces in which the student would insert his answer or the absence of such an answer column with the result that the student would write his solution directly on the test paper itself and (b) the presence of work space on the test paper itself or the absence of such workspace requiring the student to do his computations on a separate sheet of scratch paper. Although some research has been done with primary and elementary school children regarding use of answer sheets (Cashen, 1969; Gaffney, 1971; and McKee, 1967), with slow learners (Clark, 1968), and with culturally disadvantaged pupils (Soloman, 1971), virtually no studies have been reported on use of answer sheet formats with junior high school children (Miller and Minor, 1963). No research was known to the writers concerning the possible relationship of test performance to the availability or lack of availability of workspace on the test sheet itself. Thus it appeared that important implications for the validity of testing procedures underlying arithmetic computational tasks might be forthcoming from an investigation of four formats of testing involving permutations of two modes of recording answers and two types of spatial provisions for working problems.

### *Methodology*

At each of the two grade levels a quasi-experimental  $2 \times 2$  factorial design was employed involving use of four intact classes (30 students per class), each of which was randomly assigned to one of the four treatments (formats). Pretest data provided by the Comprehensive Test of Basic Skills (CTBS)—Mathematics Computation, Level 3, Form Q, revealed no statistically significant differences among the means of the four classes of seventh-grade pupils,  $F(3, 116) = 1.49, p > .05$ , or among the means of the four classes of eighth-grade pupils,  $F(3, 116) = 1.01, p > .05$ . Thus, in terms of CTBS scores no systematic differences existed in the average level of computational skills of the four classes within each grade level. Hence the groups were considered comparable.

For the 30-item teacher-made test in arithmetic, which consisted of addition, subtraction, multiplication, and division problems involving integers and fractions and which was timed at 40 minutes, the four formats representing the treatments were as follows: (1) *workspace* provided on the test paper (form) and a detached *answer column* on a separate sheet for reporting answers (WS—AC), (2) *workspace* pro-

vided on the test paper and *no* detached *answer column* resulting in the examinee's having to write answers on the test paper (WS—NAC), (3) *no workspace* provided on the test paper (though scratch paper was furnished) and a detached *answer column* on a separate sheet for reporting answers (NWS—AC), and (4) *no workspace* provided on test paper (though scratch paper was furnished) and *no* detached *answer column* resulting in the examinee's having to write answers on the test paper (NWS—NAC). This design permitted a two-way analysis of variance of the resulting scores at each grade level.

### Findings

At the seventh-grade level, the means and standard deviations of the scores of the four subsamples exposed to the four treatments WS—AC, WS—NAC, NWS—AC, and NWS—NAC were, respectively, 24.20 and 4.74, 26.13 and 3.81, 22.80 and 3.92, and 24.40 and 4.15; at the eighth-grade level, the corresponding means and standard deviations were 21.07 and 4.10, 24.43 and 4.01, 19.73 and 3.81, and 23.90 and 3.89. The analyses of variance of the scores from which these statistics were derived are summarized in Table 1. It is evident that for the seventh-grade sample both main effects were statistically significant beyond the .05 level (but not at the .01 level) but that for the eighth-grade sample only the main effect pertaining to the presence or absence of a detached answer column was statistically significant (ac-

TABLE 1  
*Two-Way Analyses of Variance of Scores on the Teacher-Made Mathematics Test of Computational Skills for Seventh and Eighth-Grade Samples*

Source of Variation	Seventh Grade Sample		MS	F
	SS	df		
Workspace (WS)	73.63	1	73.63	4.23*
Answer Column (AC)	93.63	1	93.63	5.38*
Interaction	0.83	1	0.83	0.05
Within Samples	2018.27	116	17.40	—
Total	2186.36	119	185.49	
Source of Variation	Eighth Grade Sample		MS	F
	SS	df		
Workspace (WS)	26.13	1	26.13	1.67
Answer Column (AC)	425.63	1	425.63	27.22***
Interaction	4.80	1	4.80	0.31
Within Samples	1813.80	116	15.64	—
Total	2270.36	119	472.20	

\* Significant at or beyond the .05 level.

\*\*\* Significant at or beyond the .001 level.



tually beyond the .001 level). No statistically reliable interaction effects between treatment factors occurred for either sample of 120 students.

For the seventh-grade sample, the means and standard deviations of the two subgroups of 60 subjects allowed or not allowed work space (WS vs. NWS) irrespective of answer column availability were, respectively, 25.17 and 4.38 and 23.60 and 4.08. Corresponding statistics for the two subgroups at the eighth-grade level were 22.75 and 4.36 and 21.82 and 4.36. Relative to the presence or absence of a detached answer column (AC vs. NAC), irrespective of the presence or absence of workspace provided, the means and standard deviations were, respectively, 23.50 and 4.37 and 25.27 and 4.37; similarly, at the eighth-grade level, the means and standard deviations were, respectively, 20.40 and 3.98 and 24.17 and 3.92.

In the absence of any theoretical orientation that would permit directional predictions of outcomes associated with the treatments in this experiment, significance tests were nondirectional. Thus at a descriptive level but *not* at an inferential level it is apparent that for the seventh-grade sample but to a much lesser degree for the eighth-grade sample, the mean performance was higher for the group of students provided with workspace on the test paper than for that group not given such a space allowance. Again at a descriptive level, it is evident that for the group of seventh-grade students and particularly for the group of eighth-grade students *not given* the detached answer column (on a separate sheet) the means were higher than those means for the groups of students required to record their responses in a detached answer column.

### *Discussion*

The results of this study revealed significant differences in average performance on a test of arithmetic computation for both seventh- and eighth-grade students depending upon whether a detached answer column rather than the test paper itself was employed for recording item responses. For seventh-grade but not eighth-grade students a significant difference between means occurred in relation to whether workspace was or was not provided on the test form. Interpretation of the findings at a descriptive rather than an inferential level suggests that requiring students to insert the numerical entries obtained for their problem solutions on a separate answer sheet rather than having them write these numerical answers on the test paper itself could result in lower levels of test performance. Furthermore, at least for one sample of seventh-grade students, the evidence also suggests that pro-

viding workspace on the test sheet would be facilitating to students demonstrating their level of skill in arithmetic computational tasks. Thus, it would not seem unreasonable to believe that variations in testing procedures in relation to the format of answer sheets and the provision of working space on the examination booklet could affect the validity of the scores of students taking examinations in arithmetic that emphasize computational skills.

## REFERENCES

- Cashen, V. M. The use of separate answer sheets by primary age children. *Journal of Educational Measurement*, 1969, 6, 155-157.
- Clark, C. A. The use of separate answer sheets in testing slow-learning pupils. *Journal of Educational Measurement*, 1968, 5, 61-64.
- Gaffney, R. F. Use of optically scored test answer sheets with young children. *Journal of Educational Measurement*, 1971, 8, 103-106.
- McKee, L. E. Third graders learn to use machine scored answer sheets. *School Counselor*, 1967, 15, 52-53.
- Miller, I. and Minor, F. J. Influence of multiple-choice answer form design on answer marking performance. *Journal of Applied Psychology*, 1963, 47, 347-379.
- Soloman, A. The effect of answer sheet format on test performance by culturally disadvantaged fourth-grade elementary school pupils. *Journal of Educational Psychology*, 1971, 62, 121-124.



## THE CONCURRENT VALIDITY OF THE PRIMARY SELF- CONCEPT SCALE FOR A SAMPLE OF THIRD-GRADE CHILDREN

JOAN M. JENSEN AND JOAN J. MICHAEL

California State University, Long Beach

WILLIAM B. MICHAEL

University of Southern California

In addition to estimates of the reliability for the eight factor scales of the 24-item Primary Self-Concept Scale (PSCS) obtained from two administrations to a sample of 83 children in the third grade, concurrent validity coefficients of the eight scales were determined relative to the same eight factors on the Teacher Questionnaire (TQ) designed to reflect teachers' perceptions of children's behaviors in these eight factor categories. Scores on four scales of the PSCS on its first administration and on three scales on its second administration yielded statistically significant validity ( $\phi$ ) coefficients with scores on corresponding factor categories of the TQ.

FOR a sample of 62 children ranging in age from 8 years 9 months to 9 years 2 months from the third grade of one school located in a low-middle socio-economic area of southern California, the purpose of this study was to determine the degree of concurrent validity between scores on each of eight factor scales of the 24-item Primary Self-Concept Scale (PSCS) (Müller and Leonetti, 1972) relative to scores on eight corresponding factor categories appearing on a Teacher Questionnaire (TQ) prepared by the first cited author of this paper. Thus the relationship between self-report scores of children on the eight scales of the PSCS with teacher observations on matching categories in the TQ was sought as a means of obtaining some evidence of the validity of the PSCS.

*Description of the PSCS and TQ*

Designed to provide an economic procedure for evaluation of several characteristics of self-concept relevant to school success, the PSCS was specifically constructed for use with children of Spanish or Mexican families in the Southwest, although Muller and Leonetti (1972) reported that the test was appropriate for use with children from the Anglo culture. On the basis of a factor analysis of a preliminary form, Muller and Leonetti redesigned the instrument by retaining items with high factor loadings and by adding 10 new items. Thus they had a 24-item test intended to reflect eight factors (although Item 23 dealing with two shades of skin color was omitted from scoring in this investigation). In pictorial form each item depicts at least one child in a positive role and at least one child in a negative role with the exception of Item 23. After being told a simple descriptive story about each of the illustrations, a child is instructed to draw a circle around the person that is most like himself. In Table 1, each of the eight intended factors, a brief description of the factor, the numerical designations of the items corresponding to the factor, and a brief statement of the situation portrayed by the item are presented.

The TQ contains 12 pairs of opposite-meaning words or phrases that were selected from words or phrases employed in the description of the eight factors of the PSCS set forth in Table 1. These contrasting pairs of words or phrases were subsumed under a factor category heading and were placed side by side in a left-right direction. Above each of these two words or expressions, a line about 1.25 inches long was placed so that a teacher could insert a check mark in the blank formed by the line above the word or expression which was perceived to represent behavior more like than unlike that of the child. The eight factor categories and the associated bipolar words or phrases (expressions) and corresponding item numbers may be summarized as follows: (1) *Relationship with peers I*: aggressive—cooperative (Item 1); (2) *Relationship with peers II*: ostracized—accepted (Item 2); (3) *Intellectual self-image*: likes school—dislikes school (Item 3); (4) *Helpfulness*: helped by others—not helped by others (Item 4) and helps others—does not help others (Item 5); (5) *Physiological self-image*: small—large (Item 6) and weak—strong (Item 7); (6) *Adult acceptance*: teacher acceptance—teacher rejection (Item 8) and parental acceptance—parental rejection (Item 9); (7) *Emotional self*: happy—sad (Item 10) and angry—not angry (Item 11); and (8) *Tasks undertaken* (success level): successful—unsuccessful (Item 12).

In the PSCS each item was scored one point for a socially acceptable answer and zero points for a socially undesirable response as



TABLE 1  
*Primary Self-Concept Factors and Associated Item Descriptions*

Factor Number and Description	Item Number and Description
1. <i>Peer aggressiveness or cooperation</i> Child's view of himself in sharing and cooperating with peers.	11. Loving dog—hitting dog 20. Sharing candy—fighting 22. Fighting—sharing toy
2. <i>Peer ostracism or acceptance</i> Child's view of his acceptance by his fellow students.	16. In peer group—out of peer group 18. Playing with others—playing alone 24. Playing together—playing alone
3. <i>Intellectual self-image</i> Child's view of himself as a student, and his like or dislike of school.	2. Reading—bothering child reading 4. Looking out window—reading 10. Doing school work well—not doing well 19. Reading—playing in school room
4. <i>Helpfulness</i> Assesses child's role as helper or helpee as seen by himself.	1. Helping—being helped 8. Climbing—helping child climb 9. Riding in wagon—pushing wagon
5. <i>Physiological self</i> Child's view of his physical self: large or small, strong or weak, dark-skinned or light-skinned	13. Small child—large child 15. Small child playing—large child playing 21. Strong—weak 23.* Dark child—light child
6. <i>Adult acceptance or rejection</i> Child's view of parents and teachers as accepting or rejecting him.	3. Helping mother—running from mother 5. Spanked by mother—loved by mother 14. Liked by teacher—scolded by teacher
7. <i>Emotional self</i> Laughing or crying, happy or sad, angry or not angry	7. Crying—laughing 12. Sad—happy
8. <i>Success or non-success</i> Child's view of himself as to success at task-oriented pursuits.	6. Building house—not able to build house 17. Fixing puzzle—not able to fix puzzle

\* Unscored item excluded in this research investigation.

predetermined by the test authors. For each factor scale the score was simply the number of points earned, which could vary from zero to four depending on the number of items. For a given factor Muller and Leonetti considered a score of two or more to be socially desirable. A similar procedure was followed for the PQ.

### *Data Analyses*

Phi coefficients were calculated between scores on corresponding factors of the PSCS and TQ measures after the narrow distribution of

scores was dichotomized as close to the median value as possible on each variable. Phi coefficients were evaluated for significance through use of the chi-square statistic.

Since the PSCS was given a second time two months after the first administration and that of the TQ, it was possible to obtain two sets of validity coefficients for the TQ as well as test-retest reliability estimates of the eight factor scales of the TSCS. Although on the first administration of the TSCS 101 children (54 boys and 47 girls) in the classes of four third-grade teachers participated, only 83 were present for the retest and thus available for providing data necessary to obtaining test-retest reliability estimates of the factor scores. However, since only three of the four teachers evaluated children on the TQ who had been present for the initial test and the retest of the PSCS, the sample size decreased to only 62 for the validity coefficients.

### *Findings*

In Table 2 the test-retest (phi) coefficients for the eight factor scales of the PSCS as well as the two sets of validity (phi) coefficients are cited along with levels of significance. The results may be summarized as follows:

1. Although the first six of the PSCS factor scales showed statistically significant reliability estimates varying from .25 to .69, four of the scales (1, 2, 5, and 6) of the initial administration yielded statistically significant validity coefficients ranging from .24 to .57, and three of the scales (2, 5, and 6) of the second administration furnished statistically significant validity coefficients within the span of .29 to .33.
2. As would be anticipated, PSCS factor scales showing nonsignificant or relatively low reliability estimates failed to correlate significantly with the PQ criterion measure.

### *Discussion*

Although several of the PSCS scales yielded statistically significant initial test-retest correlations, the magnitudes of these correlations were relatively low as could be anticipated for scales of only 2, 3, or 4 items. Furthermore, the phi coefficient represents a gross estimate of reliability. The size of a phi coefficient is very much influenced by the score point at which a dichotomy is formed as well as by the corresponding proportions of individuals above or below the point of dichotomy on each of the two variables being correlated. Similarly, the validity coefficients as phi values were subject to the same limitations

TABLE 2  
*Reliability Estimates (Initial Test-Retest Phi Correlations) of Each of the Eight Primary Self-Concept Scale (PSCS) Factors and Validity Coefficients (Phi Correlations) of Each of the Eight PSCS Scales with Its Corresponding Factor Category of the Teacher Questionnaire (TQ)*

Factor Number	Brief Factor Description	PSCS Item Number	TQ Category Number for Bipolar Expression	Reliability Estimates for PSCS ( $\phi$ ) Initial Test-Retest Correlations ( $N = 83$ )	Concurrent Validity Coefficients ( $\phi$ )	
					PSCS Initial Test with TQ ( $N = 62$ )	PSCS Retest with TQ ( $N = 62$ )
1.	Peer aggressiveness or cooperation	11, 20, 22	1	.69**	.27*	.20
2.	Peer ostracism or acceptance	16, 18, 24	2	.55**	.49**	.29*
3.	Intellectual self-image	2, 4, 10, 19	3	.25*	.16	.14
4.	Helpfulness	1, 8, 9	4, 5	.34**	-.19	-.02
5.	Physiological self	13, 15, 21, 23*	6, 7	.29*	.24*	.33**
6.	Adult acceptance or rejection	3, 5, 14	8, 9	.57**	.57**	.30*
7.	Emotional self	7, 12	10, 11	.00	.00	-.05
8.	Success or non-success (in task-oriented pursuits)	6, 17	12	-.02	-.04	-.04

\* Item excluded in this investigation.

\* Significant at or beyond the .05 level.

\*\* Significant at or beyond the .01 level.

and dependent upon the reliabilities of the PSCS factor scores as well as upon the reliabilities of the TQ factor category scores, which could not be estimated.

Little is also known of the difficulties that the children might have experienced not only in understanding the test tasks but also in selecting their choices in the pictorial format of the items of the PSCS. Moreover, the existence of possible response sets in the children such as acquiescence or social desirability, particularly on the retest, could have distorted the outcomes. Furthermore, the lack of teacher information about child behaviors sought in certain score categories or the lack of clarity or uniformity of meaning intended for a given category, which might have occurred from one teacher to another on both the PSCS and TQ, could be expected to attenuate the coefficients of correlation.

Although the PSCS showed some initial promise, caution should be exercised in its use. Additional research in developmental efforts in refining the items, in possibly adding one or two more items to each scale, and in correlating the scores on revised scales with behaviorally oriented criterion measures might be anticipated to improve both the reliability and validity of the PSCS factor scales.

#### REFERENCE

- Muller, D. G. and Leonetti, R. *Primary self-concept scale, study for the National Consortia for Bilingual Education*. Washington, D.C.: Office of Education (DHEW), 1972. (ERIC Document Reproduction Service No. ED 062 847)

## THE RELATIONSHIP BETWEEN SELF-ESTEEM AND ANXIETY IN GRADES FOUR THROUGH EIGHT

MARGARET A. MANY

Western Illinois University

WESLEY A. MANY

Northern Illinois University

This study examined the relationships between two measures of self-esteem and each of two measures of general anxiety and test anxiety in a sample of 4,367 pupils, grades four through eight. Coopersmith's Self-Esteem Inventory (SEI) was used to assess self-esteem. Sarason's General Anxiety Scale for Children (GASC) and Test Anxiety Scale for Children (TASC) were employed to measure anxiety. There were statistically significant negative correlations between the measure of self-esteem and each of the measures of general anxiety and test anxiety when scores were analyzed by total group, grade level, and sex. Although these correlations tended to be low to moderate ( $-.24$  to  $-.42$ ), they were consistent in suggesting a negative relationship between a measurable construct of self-esteem with each of the corresponding constructs of general and test anxiety. The implications tend to support the possibility of reducing anxiety in elementary and junior high school age pupils by enhancing the way in which they see themselves.

RESEARCH findings generally indicate that persons with high self-esteem are happier and more effective in meeting societal demands than are persons with low self-esteem. Findings further point to the undesirable consequences that can accrue as a result of extreme anxiety within the individual. Coopersmith (1967) has suggested that this anxiety may occur when the individual expects to be or actually is rejected by himself or by others.

In a study involving fourth, fifth, and sixth grade pupils, Lipsitt



(1958) found that children with poor self-concept were significantly more anxious than were children with good self-concepts. From his study with college females Mitchell (1959) concluded that the better the self-concept, the less anxiety evidenced. Imbler (1968) obtained a significant negative correlation between anxiety and positive self-concept. In a related study involving a comparison between high and low self-esteem subjects Lampl (1968) observed that the low self-esteem subjects were higher in anxiety than were the high self-esteem subjects. Similarly, studies by Van Buskirk (1961), Wittrock and Husek (1962), Coopersmith (1967), and Ausubel and Robinson (1969) have provided evidence that a negative relationship between level of anxiety and favorableness of self-concept or self-esteem appears to exist.

The purpose of this study was to examine the relationship between a measure of self-esteem and (a) a measure of general anxiety and (b) one of test anxiety in a large population and in subpopulations of elementary school children in grades four through eight. The measures employed were Coopersmith's Self-Esteem Inventory (SEI) and Sarason, Lighthall, Davidson, Waite, and Ruebush's (1960) General Anxiety Scale for Children (GASC) and Test Anxiety Scale for Children (TASC). To the writers' knowledge a study of the relationship between Coopersmith's Self-Esteem Inventory and Sarason's anxiety scales had not been previously conducted. This study was intended to provide further data concerning the degree of relationships between a construct of self-esteem and each of two constructs of anxiety in terms of their being represented operationally by these particular instruments that have gained increased acceptance by educators.

### *Subjects*

The subjects consisted of 4,367 pupils from public schools in grades four through eight in East Aurora and in Wheaton, Illinois. This sample included all students from these grade levels in the two communities, except those for whom both test scores were unavailable and for those students who had an anxiety lie score of five or lower. The diversity of these two communities, when combined, affords a broad range of socio-economic status as well as representation from different ethnic and racial minority groups.

### *Procedure*

The SEI, GASC, and TASC were administered to all children in grades four through eight of the participating schools. The scales were distributed to each teacher responsible for a grade or class. Both scales

TABLE 1  
*Correlation between Scores on SEI Scale and Scores on GASC Scale  
 and between Scores on SEI Scale and Those on TASC Scale*

Paired Scales	N	r
SEI and GASC	4,367	-.280*
SEI and TASC	4,367	-.381*

\* Significant beyond the .001 level.

were administered according to standard directions accompanying the scales. The two anxiety scales, GASC and TASC, were given to the pupils at one testing period and the SEI at another. The scales were read to all groups. The testing was accomplished during the second week of May. It was assumed that by this time in the school year anxiety a student might feel as a result of being in a new school, new class, or new program would be greatly diminished.

After being assembled in an appropriate form the data were checked for accuracy by the research directors. Pearson product moment coefficients of correlation were calculated between sets of scores and the data were analyzed by total group score, by grade level, and by sex.

### Results

As is evident in Table 1, significant negative relationships were found between the measure of self-esteem and each of those of general anxiety and test anxiety for the total population of children.

When analyzed by grade level, similar statistically significant negative relationships were found. As is apparent in Table 2, the highest relationships were found among the sixth grade pupils.

TABLE 2  
*Correlation between SEI Scores and (a) GASC Scores and  
 (b) TASC Scores within Each of Five Grade Levels*

Paired Scales	Grade Level	N	r <sup>a</sup>
SEI and GASC	4	824	-.243
	5	816	-.287
	6	960	-.318
	7	854	-.299
	8	913	-.245
SEI and TASC	4	824	-.388
	5	816	-.399
	6	960	-.424
	7	854	-.366
	8	913	-.319

\* All correlation coefficients significant beyond the .001 level.

(1958) found that children with poor self-concept were significantly more anxious than were children with good self-concepts. From his study with college females Mitchell (1959) concluded that the better the self-concept, the less anxiety evidenced. Imbler (1968) obtained a significant negative correlation between anxiety and positive self-concept. In a related study involving a comparison between high and low self-esteem subjects Lampl (1968) observed that the low self-esteem subjects were higher in anxiety than were the high self-esteem subjects. Similarly, studies by Van Buskirk (1961), Wittrock and Husek (1962), Coopersmith (1967), and Ausubel and Robinson (1969) have provided evidence that a negative relationship between level of anxiety and favorableness of self-concept or self-esteem appears to exist.

The purpose of this study was to examine the relationship between a measure of self-esteem and (a) a measure of general anxiety and (b) one of test anxiety in a large population and in subpopulations of elementary school children in grades four through eight. The measures employed were Coopersmith's Self-Esteem Inventory (SEI) and Sarason, Lighthall, Davidson, Waite, and Ruebush's (1960) General Anxiety Scale for Children (GASC) and Test Anxiety Scale for Children (TASC). To the writers' knowledge a study of the relationship between Coopersmith's Self-Esteem Inventory and Sarason's anxiety scales had not been previously conducted. This study was intended to provide further data concerning the degree of relationships between a construct of self-esteem and each of two constructs of anxiety in terms of their being represented operationally by these particular instruments that have gained increased acceptance by educators.

### *Subjects*

The subjects consisted of 4,367 pupils from public schools in grades four through eight in East Aurora and in Wheaton, Illinois. This sample included all students from these grade levels in the two communities, except those for whom both test scores were unavailable and for those students who had an anxiety lie score of five or lower. The diversity of these two communities, when combined, affords a broad range of socio-economic status as well as representation from different ethnic and racial minority groups.

### *Procedure*

The SEI, GASC, and TASC were administered to all children in grades four through eight of the participating schools. The scales were distributed to each teacher responsible for a grade or class. Both scales

TABLE 1

*Correlation between Scores on SEI Scale and Scores on GASC Scale  
and between Scores on SEI Scale and Those on TASC Scale*

Paired Scales	N	r
SEI and GASC	4,367	-.280*
SEI and TASC	4,367	-.381*

\* Significant beyond the .001 level.

were administered according to standard directions accompanying the scales. The two anxiety scales, GASC and TASC, were given to the pupils at one testing period and the SEI at another. The scales were read to all groups. The testing was accomplished during the second week of May. It was assumed that by this time in the school year anxiety a student might feel as a result of being in a new school, new class, or new program would be greatly diminished.

After being assembled in an appropriate form the data were checked for accuracy by the research directors. Pearson product moment coefficients of correlation were calculated between sets of scores and the data were analyzed by total group score, by grade level, and by sex.

### Results

As is evident in Table 1, significant negative relationships were found between the measure of self-esteem and each of those of general anxiety and test anxiety for the total population of children.

When analyzed by grade level, similar statistically significant negative relationships were found. As is apparent in Table 2, the highest relationships were found among the sixth grade pupils.

TABLE 2

*Correlation between SEI Scores and (a) GASC Scores and  
(b) TASC Scores within Each of Five Grade Levels*

Paired Scales	Grade Level	N	r <sup>a</sup>
SEI and GASC	4	824	-.243
	5	816	-.287
	6	960	-.318
	7	854	-.299
	8	913	-.245
SEI and TASC	4	824	-.388
	5	816	-.399
	6	960	-.424
	7	854	-.366
	8	913	-.319

<sup>a</sup> All correlation coefficients significant beyond the .001 level.

TABLE 3  
*Correlation between SEI Scores and (a) GASC Scores and  
 (b) TASC Scores by Sex*

Paired Scales	Sex	N	r <sup>a</sup>
SEI and GASC	Male	1,997	-.289
	Female	2,320	-.272
SEI and TASC	Male	1,997	-.377
	Female	2,320	-.381

<sup>a</sup> All correlation coefficients significant beyond the .001 level

### *Conclusions and Discussion*

There were statistically significant negative correlations between a measure of self-esteem and each of the measures of general anxiety and test anxiety when scores were analyzed by total group, by grade level, and by sex. Although these correlations tended to be low to moderate (-.24 to -.42), they were consistent in suggesting a negative relationship between a measurable construct of self-esteem with each of the corresponding constructs of general anxiety and test anxiety.

The findings of this study based on the use of particular instruments selected generally supported the outcomes of other similar research. Although a correlational study of this nature does not deal with cause and effect relationships, the implications tend to support the possibility of reducing anxiety in elementary and junior high school age children by enhancing the way in which these children see themselves. It would appear reasonable to suggest that efforts to provide opportunities for successful achievement should be undertaken in an environment that reinforces the adequacy and worthiness of the individual student.

### REFERENCES

- Ausubel, D. P. and Robinson, F. G. *School learning: An introduction to educational psychology*. New York: Holt, Rinehart and Winston, 1969.
- Coopersmith, S. *The antecedents of self-esteem*. San Francisco: W. H. Freeman, 1967.
- Imbler, I. I. The effects of participation training on closed-mindedness, anxiety, and self-concept (Doctoral dissertation, Indiana University, 1967). *Dissertation Abstracts*, 1968, 28, 3451A. (University Microfilms No. 68-2305)
- Lampl, M. Defensiveness, dogmatism, and self-esteem (Doctoral dissertation, Yeshiva University, 1968). *Dissertation Abstracts*, 1968, 29, 2194B. (University Microfilms No. 68-17, 171)
- Lipsitt, L. P. A self-concept scale for children and its relationship to



- children's form of the manifest anxiety scale. *Child Development*, 1958, 29, 463-472.
- Mitchell, J. V., Jr. Goal-setting behavior as a function of self-acceptance, over- and underachievement, and related personality variables. *Journal of Educational Psychology*, 1959, 50, 93-104.
- Sarason, S. B., Davidson, K. S., Lighthall, F. F., Waite, R. R., and Ruebush, B. K. *Anxiety in elementary school children*. New York: Wiley, 1960.
- Van Buskirk, C. Performance on complex reasoning tasks as a function of anxiety. *Journal of Abnormal and Social Psychology*, 1961, 62, 201-209.
- Wittrock, M. C. and Husek, T. R. Effect of anxiety upon retention of verbal learning. *Psychological Reports*, 1962, 10, 78.



## THE PREDICTIVE VALIDITY OF THE TEST OF AUDITORY PERCEPTION

CARL R. SCHMIDT

Model Learning  
Disabilities Systems  
State College, Pennsylvania

DAVID A. SABATINO

Northern Illinois  
University

GLEN G. FOSTER

Model Learning  
Disabilities Systems  
State College, Pennsylvania

Eighty-four second grade children were administered the Lorge-Thorndike Test of Cognitive Abilities (LITCA), the Test of Auditory Perception (TAP) and, five months later, the Metropolitan Achievement Test (MAT). Pearson product-moment correlation coefficients and an analysis using stepwise regression indicated that performance on the TAP subtests was significantly correlated with most MAT subtest scores, especially the phonically-oriented MAT Word Analysis subtest. With the common TAP-LITCA variance held constant, all correlation coefficients which had previously been statistically significant became non-significant except those between the TAP subtests and the MAT Word Analysis subtest. It was concluded that performance in only those aspects of an academic program most directly related to auditory perception can be predicted using the TAP.

THE Test of Auditory Perception (TAP) (Sabatino and Foster, 1974) is an experimental device designed to facilitate prescription of educational programs for learning disordered children. It has been

well established that auditory perception is related to academic achievement, usually measured by reading ability (Benger, 1968; Robinson and Hanson, 1968; Wepman, 1960). It has, therefore, been recommended that information relating to auditory perception be used in formulating educational programs (Johnson and Myklebust, 1967; Wepman, 1960). Empirical evidence, however, does not support the efficacy of this practice (Waugh, 1973; Ysseldyke, 1973). Ysseldyke (1973) suggested that at least part of the difficulty in prescribing academic programs on the basis of a child's auditory or visual perceptual abilities is the "lack of reliable and valid devices which may be used to identify behavioral (ability) strengths and weaknesses in children" (p. 26).

The purpose of this study was to establish the predictive validity of the TAP relative to academic achievement, and to verify that auditory perception as measured by the TAP is a unique ability construct, i.e., that it is not primarily a measure of general intelligence.

### *Procedure*

#### *Subjects*

The pupils in an entire second grade ( $n = 84$ ) attending a rural elementary school participated in this study.

#### *Experimental Instrument*

The Test of Auditory Perception (TAP) consists of three subtests: (1) Phoneme Discrimination requiring differentiation among three similar sounding nonsense syllables, (2) Word Recognition involving identification of a legitimate word among a set of similar sounding nonsense words, (3) Sequencing necessitating the matching of a series of nonsense syllables to one of three multisyllabic nonsense words. KR-20 reliabilities on the subtests range from  $r = .896$  to  $r = .740$  depending on subtest and age level.

### *Method*

All subjects were administered the Lorge-Thorndike Test of Cognitive Abilities (LTTCA) (Lorge, Thorndike, and Hagen, 1964) and the TAP. Five months later the battery of Metropolitan Achievement Test (MAT) (Durost, Bixler, Wrightstone, Prescott and Balow, 1971) was administered.

### *Results*

Each TAP and MAT subtest was treated separately for purposes of analysis. To investigate the simple relationships among the individual

subtests, Pearson product moment correlations ( $r$ ) were computed. These results appear in Table 1. All TAP subtests correlated significantly ( $p < .05$ ) with all MAT subtests except the TAP Discrimination versus the MAT reading subtest ( $r = .19$ ) and the TAP Sequencing subtest versus the MAT Word Knowledge subtest ( $r = .17$ ).

To investigate the possibility that the relationship between the TAP subtests and the MAT subtest was due to overlap with a general intelligence factor, partial correlation coefficients (Partial  $r$ ) were computed between each TAP subtest and each MAT subtest holding LTTCA scores constant. These results also appear in Table 1.

All correlations between the TAP subtests and MAT Word Knowledge, Reading and Mathematics were nonsignificant after LTTCA scores were partialled out. Those correlations between the TAP subtests and the MAT Word Analysis subtest, however, remained significant ( $p < .01$ ).

To investigate the possibility that a weighted combination of TAP subtests might predict some MAT subtests more accurately than would any individual TAP subtest, a stepwise regression was computed using TAP subtests as predictors and each of the MAT subtests as a dependent variable. The results appear in Table 2. The accuracy with which the individual TAP subtests predicted performance on the MAT subtests was increased by using a weighted combination of TAP subtests. This increment in validity was especially notable in the case of the MAT Word Analysis subtest.

### Discussion

The MAT Word Knowledge subtest had low correlation coefficients with all of the TAP subtests. The corresponding coefficients of the TAP subtests were also low relative to the MAT Reading subtest. Even though the Mathematics subtest had modest correlations with the

TABLE 1  
Zero-Order Correlation Coefficients between TAP Subtests and MAT Subtests and Corresponding Partial Correlation Coefficients Holding LTTCA Scores Constant

MAT Subtests	Discrimination		TAP Subtests Recognition		Sequencing	
	$r$	Partial $r$	$r$	Partial $r$	$r$	Partial $r$
Word Knowledge	.23*	.13	.27*	.12	.17	.05
Word Analysis	.31**	.26**	.34**	.26**	.32**	.26**
Reading	.19	.08	.25*	.08	.25*	.14
Mathematics	.26*	.14	.29**	.09	.30**	.07

\*  $p < .05$ .

\*\*  $p < .01$ .



TABLE 2  
*Stepwise Regression Analysis Using TAP Subtests as Predictors,  
 and MAT Subtests as Dependent Variables*

Dependent Variable	Independent Variable(s)	Regression Coefficient	Correlation Coefficient	Multiple Correlation Coefficient	Proportion of Explained Variance
Word	Recognition	.1196	.27*		
Knowledge	Discrimination	.1404	.23*	.31**	.11
Word	Recognition	.1078	.34**		
Analysis	Discrimination	.1319	.31**		
	Sequencing	.5081	.32**	.45***	.20
Reading	Sequencing	.6531	.25*		
	Recognition	.9655	.25*	.31**	.10
Mathematics	Recognition	.1057	.29**		
	Sequencing	.7596	.30**	.36**	.13

\*  $p < .05$ \*\*  $p < .01$ \*\*\*  $p < .001$ 

TAP subtests, these correlations dropped to nearly zero after partialling out variance associated with the LTTC A. There are several possible explanations for these results. The low correlations between the TAP subtests and the MAT Word Knowledge and Reading subtests might indicate that these tasks, as measured, do not relate significantly to auditory perception, as tested. The marked decrease in correlation coefficients between the TAP subtests and the three MAT subtests previously mentioned, which was due to partialling out variance associated with intelligence test scores, might indicate either that these tasks require a high degree of cognitive mediation which may mask the importance of auditory perceptual functioning or that general intelligence tests measure a significant component of auditory perception.

Nevertheless, all TAP subtests did produce a zero-order correlation coefficient above .30 ( $p < .01$ ) with the MAT Word Analysis subtest. Further the combined TAP subtests produced a multiple correlation coefficient of .45 ( $p < .001$ ) and the TAP Word Analysis correlation was only slightly reduced by partialling out variance in common with LTTC A scores. The Word Analysis subtest appears to be the MAT subtest most directly related to auditory perception, which may account for the higher, more stable coefficients.

It can be concluded from the preceding data analysis that only those aspects of academic achievement which can be directly associated with auditory functioning, at a very basic level, can be accurately predicted from the TAP subtests. Therefore the TAP may prove useful in determining which children will probably function adequately in an aca-

demic program stressing auditory skills, e.g., phonic reading, and which children may have undue difficulty if forced through such a program.

Perhaps the most potentially important conclusion that can be inferred from this study is that it supports the possibility that performance in specific skill areas in reading can be predicted. Global tests, such as intelligence measures, can accurately predict global results such as reading ability; but results obtained from global measures are of limited diagnostic use. If utilitarian subskills can be determined, e.g., phonic word analysis, and if researchers can devise measures which can reliably and validly predict a given child's ability in these subskill areas, it may prove a useful teaching strategy to use this information either to strengthen the area of weakness or to circumvent a wall of frustration.

### REFERENCES

- Benger, K. The relationship of preception, personality, intelligence, and grade one reading achievement. In H. K. Smith (Ed.), *Perception and Reading*. Newark, Delaware: International Reading Association, 1968.
- Durost, W., Bixler, H., Wrightstone, J., Prescott, G., and Balow, I. *Metropolitan Achievement Test*. New York: Harcourt, Brace and Jovanovich, 1970.
- Johnson, D. and Myklebust, H. *Learning disabilities: Educational principles and practices*. New York: Grune and Stratton, 1967.
- Lorge, I., Thorndike, R., and Hagen, E. *Lorge Thorndike Test of Cognitive Abilities*. New York: Houghton Mifflin Co., 1964.
- Orlando, C. Review of the reading research in special education. In I. Mann and D. Sabatino (Eds.), *The first review of special education*. Vol. I. Philadelphia: Buttonwood Farms, Inc., 1973.
- Robinson, H. and Hanson, E. Reliability of measures of reading achievement. *The Reading Teacher*, 1968, 21, 307-313.
- Sabatino, D. and Foster, G. Personal communication, 1973, 1974.
- Waugh, R. Relationship between modality preference and performance. *Exceptional Children*, 1973, 39, 465-469.
- Wepman, J. Auditory discrimination, speech, and reading. *Elementary School Journal*, 1960, 60, 325-333.
- Ysseldyke, J. Diagnostic-prescriptive teaching: The search for aptitude-treatment interactions. In I. Mann and D. Sabatino (Eds.), *The first review of special education*. Vol. I. Philadelphia: Buttonwood Farms, Inc., 1973.



## THE VALIDITY OF THE SRA ACHIEVEMENT SERIES, MULTILEVEL EDITION: READING, LANGUAGE ARTS, AND ARITHMETIC SUBTESTS FOR MINORITY AND NON-MINORITY GROUP FOURTH GRADE PUPILS

EDWARD B. TOKAR AND FREDERICK STOFFLET

Norfolk Public Schools

Significant positive correlations, with one exception, between ratings of fourth grade pupils by teachers on a three point scale and scores on the Reading, Language Arts, and Arithmetic subtests of the SRA Achievement Series, Multilevel Edition, Blue Form E suggested that these subtests would be valid measures of group academic achievement by both minority and non-minority children. A non-significant correlation between SRA reading and teacher's rating of minority group pupil reading achievement suggested a need for further investigation. Correlations between SRA subtest scores and teacher's ratings ranged from .20 to .57 for minority group pupils and from .46 to .55 for non-minority group pupils.

THE purpose of this study was to investigate the validity of the Reading, Language Arts, and Arithmetic subtests of the SRA Achievement Series, Multilevel Edition, Blue Form E. Of particular interest was the SRA's differential validity for minority and non-minority group fourth grade pupils.

To facilitate the analysis, a pupil's raw score for the Reading subtest was designated as his SRA reading score (SRAR). Similarly, a pupil's raw scores for the Language Arts and Arithmetic subtests were designated as his SRA Language Arts (SRALA) and SRA Arithmetic (SRAA) scores, respectively.

### *Subjects*

A total of 92 fourth grade pupils, within five elementary schools of the Norfolk Public Schools, participated in the investigation. The

TABLE 1  
*Intercorrelations of Predictors (SR 4 Subtests) and Criterion (Teacher's Ratings) Variables for Non-minority Pupils (above Diagonal) and Minority Pupils (below Diagonal) with Means ( $\bar{X}$ ) and Standard Deviations (SD)*

Variables	Correlations <sup>a</sup>						Minority Pupils (N = 55)		Non-minority Pupils (N = 37)	
	1	2	3	4	5	6	$\bar{X}$	SD	$\bar{X}$	SD
1 SRAR	—	.69	.57	.46	.38	.26	24.80	10.18	37.84	14.40
2 SRALA	.69	—	.30	.52	.55	.39	24.71	8.31	31.24	11.37
3 SRAA	.48	.56	—	.54	.43	.49	19.64	7.04	25.24	9.97
4 TRR	.20	.38	.41	—	.66	.44	10.07	2.68	11.11	2.95
5 TRLA	.23	.43	.52	.82	—	.60	7.51	2.25	8.22	2.09
6 TRA	.27	.51	.57	.49	.61	—	12.13	4.52	15.24	4.93

<sup>a</sup> Coefficients above the diagonal equal to or exceeding .33 are significant at the .05 level.

<sup>b</sup> Coefficients below the diagonal equal to or exceeding .28 are significant at the .05 level.



minority group was composed of 55 black pupils whereas the non-minority group was composed of 37 nonblack pupils.

### *Criteria*

The yearly average of the teacher's ratings of each pupil's performance served as the criteria. Teachers assigned ratings corresponding to a judgment of "progressing slowly," "progressing" or a "progressing rapidly." These levels were assigned the quantities 1, 2, and 3, respectively. Ratings in the factors of understanding, reading smoothly, attacking new words, building vocabulary, and reading independently were summed and designated as the Teacher's Rating of Reading (TRR). Similarly, ratings in the categories of spelling needed words, spelling assigned words, punctuation, capitalization and expressing ideas were summed and designated as the Teacher's Rating of Language Arts (TRLA). Furthermore, the sum of teacher's ratings of sets, place value, division, measurement, geometry, and fractions were designated as the criterion Teacher's Rating of Arithmetic (TRA).

### *Results and Discussion*

The SRA tests were administered in October of 1973. The means, standard deviations and intercorrelations of the SRAR, TRR, SRALA, TRLA, SRAA, and TRA for non-minority and minority pupils are shown in Table I.

All SRA subtest scores, with one exception, were significantly related to teacher's ratings for both non-minority and minority pupils beyond the .05 level of significance. The results would suggest that Language Arts and Arithmetic subtests of the SRA Achievement Series, Multilevel Edition, Blue Form E., are valid measures of fourth grade minority and non-minority group pupil achievement. However, the validity of the Reading subtest, for minority group pupils, should be re-examined.



## A PROFILE OF THE VALIDITY OF POSTGRADUATE PSYCHOLOGY EXAMINATIONS IN PAKISTAN IN TERMS OF THEIR CONGRUENCE WITH EDUCATIONAL OBJECTIVES

Z. A. ANSARI

University of Peshawar  
Peshawar, Pakistan

Within the framework provided by the six process-oriented cognitive categories in the *Taxonomy of Educational Objectives* by Bloom and his co-workers the essay-type items of postgraduate achievement examinations in psychology for students at the University of Peshawar were classified, and the relative frequencies of the items were compared with those for psychology examinations administered at a British university. Although the categories of Knowledge and Comprehension were amply represented, the dimensions of Analysis and Synthesis would appear to have been very much underrepresented at the University of Peshawar. Thus possible sources of invalidity in the achievement examinations were identified if the behaviors of the examinees both as students and as future psychologists demand abilities of Analysis and Synthesis in research endeavors, in report writing, and in oral communication as required in teaching or public service.

THE publication of the *Taxonomy of Educational Objectives* (Bloom, Engelhart, Furst, Hill, and Krathwohl, 1956) has provided a framework for evaluation of achievement tests in terms of educational objectives. Standardized achievement tests as well as teacher-made tests at all educational levels have been evaluated through using the *Taxonomy* or its variants as a model (Ansari, 1971; Bloom, 1959; McGuire, 1963; Yorkshire Regional Examinations Board, 1968).

The purpose of the present investigation was to evaluate the validity of achievement tests being given to MA/MSc Psychology students at the University of Peshawar. Since these examination papers are set by

senior teachers of psychology from other universities of Pakistan, these tests can be taken as a fair sample of the type of examinations being given to the postgraduate students of psychology in Pakistan.

### *Material*

The MA/MSc psychology examination in the University of Peshawar consists of six theory papers and of some laboratory work, evenly divided between two annual examinations. Only the theory examinations were included in this investigation. According to the custom of this country, each theory paper consists of ten or eleven essay-type questions; sometimes the last one is a short-notes question.

Since 1965, when the University of Peshawar held the first MA/MSc examinations in psychology, up to 1973, in all 534 essay-type questions were asked. Since this number is a very large one, only half of these items were used for the purpose of analysis. However, each of the three raters missed evaluating some items, with the result that the judgments by all the three raters were available for 172 items only.

### *Raters*

The raters were three students of MA Final Psychology class, who had had Psychological Testing as their major subject. The main points of Bloom's *Taxonomy* were explained to the raters, and they were provided with a copy of the main categories of classification. The ratings were done independently. The measure of agreement used was the classification of an item in the same category by at least two raters.

### *Findings and Discussion*

Out of 172 items for which the ratings by all the three raters were available, 22 items were placed in the same category by all the three raters ( $p = .005$ ), whereas on 92 items there was agreement among two out of three raters ( $p = .069$ ). In all total or partial agreement was reached on 114 items (66%). The extent of agreement is similar to what was found in an earlier investigation in which essay-type tests were used (Ansari, 1971).

The profile of the validity of examinations in terms of educational objectives is presented in Table 1. The results show that the main emphasis of Peshawar University examinations has been on cognitive objectives of the lower categories. Two of the lowest categories, Knowledge and Comprehension, account for about two-thirds of the

TABLE I

*Percentages of Items in Postgraduate Examinations of Peshawar and Glasgow Universities Assigned to Each of Six Categories of Educational Objectives*

	Peshawar	Glasgow
1. Knowledge	36	26
2. Comprehension	28	18
3. Application	14	9
4. Analysis	5	21
5. Synthesis	3	3
6. Evaluation	14	23
Totals	100	100

examination questions. The four higher categories share the rest one-third of the questions.

In an earlier study the psychology examinations of a British university (Glasgow) were analyzed in a similar manner (Ansari, 1971). The two analyses show interesting similarities as well as differences. The British university examinations also placed a heavy emphasis on lower cognitive objectives, but not to the same extent as in Pakistan. There is a greater emphasis in the British university than in Pakistan on questions requiring Analysis and Evaluation. Surprisingly, in both the universities questions of Synthesis are rather few. This observation is particularly noteworthy because those who favor the use of essay-type questions maintain that "... abilities to select, relate and organize, to create essentially new patterns" can be appraised better by the essay-type tests (Thorndike and Hagen, 1955). In spite of this potential, it seems that the essay-type tests are not being used to measure these abilities, particularly in Pakistani universities. Thus if the processes of synthesis and analysis are important in the postgraduate curricula of psychology and to the subsequent research and vocational endeavors in psychology their lack of representation in postgraduate achievement examinations may constitute significant sources of examination invalidity in need of correction or remediation.

## REFERENCES

- Ansari, Z. A. *A study of achievement tests and other psychological tests*. Unpublished Ph.D. Thesis, University of Glasgow, 1971.
- Bloom, B. S. Review of Cooperative General Culture Test. In O. K. Buros (ed.) *The fifth mental measurements yearbook*. Highland Park, N. J.: Gryphon Press, 1959, pp. 7-9.
- Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., and Krathwohl, D. R. *Taxonomy of educational objectives. Handbook I: Cognitive domain*. New York: David McKay, 1956.



- McGuire, C. Research in the process approach to the construction and analysis of medical examinations. *Twentieth yearbook of the national council on measurement in education*. East Lansing: National Council on Measurement in Education, Michigan State University, 1963, pp. 7-16.
- Thorndike, R. L. and Hagen, E. *Measurement and evaluation in psychology and education*. New York: Wiley, 1955.
- Yorkshire Regional Examinations Board. *The educational objectives of examinations with particular reference to geography*. Research Report No. 5. Harrogate, England: the Board, 1968.

## BOOK REVIEWS

Gary D. Borich (Ed.). *Evaluating Educational Programs and Products*. Englewood Cliffs, New Jersey: Educational Technology Publications, 1974. Pp. xxiii + 491, \$12.95.

The work of the evaluator is described in *Evaluating Educational Programs and Products* as consisting of the performance of three different activities: establishing perspectives, planning the evaluation, and analyzing the data. Following an introductory chapter entitled Prologue, chapters are organized according to three major themes: "Roles and Contexts," "Models and Strategies," and "Methods and Techniques for Evaluating Educational Programs and Products"; the text is concluded with a chapter entitled Epilogue. Introductory sections set the themes to the three major sections of the text.

Chapters are further classified in a reader's guide according to probable relevance of content for an individual's position in an educational system. Using this table, the reader can begin by reading the content which is likely to be of primary importance and leave the reading of other less related chapters to some later time.

Michael Scriven begins in the Prologue to enumerate some standards for use in the evaluation of educational programs and products. A Product Evaluation Profile checklist is provided for use by individuals responsible for the evaluation of educational program and product development.

Chapter 1 presents a statement by Jean W. Butman and Jerry L. Fletcher on the use of theory to provide a basis for the educational R & D process and how the developer/evaluator is constrained by practical and political influences in his/her attempt to attain these idealized states of operation.

The next four chapters are directed toward the delineation of specific activities of and/or proposed standards for the product evaluation process. Eva Baker considers the role of evaluation in the instructional development process by relating the formative evaluation needs of instructional product developers to data collection activities which can provide the required information on needs. An example of the development of an instructional product is also provided to indicate the role of formative evaluation in the developmental process and to remind the reader that the best developed plans may still result in the data to be used for making a product revision coming in after the revision has been accomplished. Chapter 3, written by Barbara J. Brandes, describes the role of formative evaluation within the devel-

opment of an instructional product designed specifically to attain affective as well as cognitive outcomes. Noting that different formative evaluation strategies are required to attain these types of outcomes, Brandes suggests that early in the instructional development process, the focus should be on attaining cognitive outcomes and that only when the cognitive part is under control should the developer turn to looking at affective goals. Alkin and Fink, in their chapter, assert that the individual who purchases instructional products is generally not given enough information on the appropriateness of the product for instructional situation or else information is not summarized in an easily digested form. Desirable requisite information items for educational product purchasers are provided along with an example of how such information has been packaged for one instructional product. Even though this chapter was oriented to people "out-in-the field," the concerns stated coincide with those presented by Scriven in the Prologue. Similar in focus to the Alkin and Fink chapter is the next chapter by Louise L. Tyler and M. Frances Klein which is concerned with specifying and discussing recommendations also categorized as to whether they are to be considered desirable, very desirable, or essential requirements in the curriculum or instructional product developmental process.

Part two in the text presents a number of different models (or views) representing procedures used for performing program and product evaluation. Wright and Hess begin in Chapter 6 by identifying the dimensions of (1) stages of evaluation, (2) audience for the evaluation, and (3) domains of evaluative criteria, so that activities to be performed in educational product evaluation process can be identified and the criteria to be achieved at each of one evaluative stage specified. The Bertram and Childers and Katz and Morgan chapters describe systems models which indicate the flow of activities and information generated through the evaluation process. While Bertram and Childers focus on specific activities and procedures in an evaluation process, the Katz and Morgan chapter is more general. Their flow model has decision points which raises questions about the extent to which goals are being attained and the need for someone to make decisions in reference to originally stated goals of the project as mediated by the context in which the program-product development operates.

The next two chapters describe quality control approaches for evaluating educational R & D efforts by Jerry P. Walker (chapter 9) and the process of process evaluation by Max Luft, Janice Lujan, and Katherine A. Bemis (chapter 10). Both of these presentations include an observation scale which has been used in such evaluation activities.

Part three contains six chapters which present the methodological considerations of formative and summative evaluation of educational programs and products. Chapter 11, *Formative Evaluation: Selecting Techniques and Procedures*, was written by James Sanders and Donald Cunningham. This chapter describes the so-called "procedures and

materials" of formative evaluation categorized according to appropriateness for each of four different stages of formative evaluation: pre-development, evaluation of objectives, interim evaluation, and product evaluation.

The remaining five chapters deal more specifically with the analysis aspect of evaluation. Borich and Drezek in chapter 12 use the correlational causal relationship methodology developed by Blalock to ascertain the validity of hypotheses relating concomitant variables and the instructional transaction resulting from the use of a new educational product. A computer program is included as an appendix to assist readers in implementing such an approach to process evaluation. Both Eichelberger and Edwards, in their chapters (13 and 14), agree with the thesis of Katz and Morgan that the evaluator needs to be aware of the environment in which the evaluation is being performed. Such information can be used to identify the level of analyses to be used in the formative and summative evaluations of educational programs and products. Both authors are very strong in the use of advanced correlational-based analysis procedures for evaluation. Poyner, in chapter 17, considers specifically the question of what is the appropriate unit of analysis of evaluation studies. He concludes that hierarchical models with classes serving as nested factors represent the analysis of procedure of choice since this procedure allows one to test the assumption of equivalency of classes. When this assumption is rejected, the class mean becomes the unit of analysis. If the assumption is not rejected, then the individual pupil can be used as the unit of analysis. The final chapter discusses some research conducted by Andrew Porter and Thomas Chibucos on the appropriateness of statistical analysis procedures for evaluation studies. They present four different types of situations and discuss the use of alternative analysis models. Their conclusion is that no one design and/or analysis strategy will work in all settings; however, the guidelines provided in the chapter will be useful to all individuals involved in the analysis of data collected to document effectiveness of a particular educational product or program.

The Epilogue written by P. Kenneth Kosmoski attempts to describe the future of instructional product development. Specifically, he feels that educational product developers must begin to provide alternative instructional modes to the printed page and that future prospective buyers of such goods will require hard evidence of the effectiveness of such projects to bring about learning.

Since instructional product and program evaluation is still a new and developing field of study, it is not possible to evaluate the contributions in this book for accuracy of content as one could do in the evaluation of a text in such a field as Statistics. In contrast, it does seem reasonable to evaluate the extent to which this book does achieve its goals.

The editor stated in the preface that the "book is not a textbook or



collection of articles, but an especially prepared guide and handbook for planners, developers, and evaluations of educational programs and products (p. vii)." In light of this statement, the reviewer was struck by the lack of continuity. Since the editor stated a long (two years) and intensive R & D cycle of review and revision was performed in the production of the text, one would expect a smooth-flowing, balanced presentation of elements of formative evaluation. Unfortunately, this is not the case. The chapters appear to have been written without any consideration or, more likely, without any knowledge of what the other authors were doing. None of the contributed chapters contain a reference to any other chapter in the text. This was surprising in view of the unavoidable overlap in context treated by the several chapters. For example, the Prologue and Epilogue chapters present some suggestions as to information needs for prospective users of instructional products, the topic covered by Alkin and Fink and yet no reference is made to any of the other's ideas on the subject. Another example was the consideration of which unit of analysis is appropriate to use in evaluation studies. This topic is considered briefly by Edwards while Poynor devotes his whole chapter to this question and neither cites the other. This reviewer feels an individual who takes on the responsibility of coordinating the gathering of information from several authors should also use the review process to inform an individual author of what the other contributors have said that is relevant to their topic.

This reviewer also feels a handbook should essentially cover the field. Thus, a reasonable approach would be to identify a number of different aspects of the instructional product-program evaluation process and then to have specific chapters written to cover each of these areas. Thus, the overlap of topics treated and wide range of specificity of content covered in the chapters was surprising in view of the claims of careful planning and selection of topics. The extreme range of specificity is denoted at one end by one chapter in the methodology section which only covers the topics of appropriateness of units of analysis and at the other extreme by a chapter which exhaustively covers measuring instruments and associated operational procedures useful in formative evaluation.

In conclusion, this reviewer feels the book will not serve as a basic reference for individuals performing formative evaluations of programs and products. However, by being selective, readers can benefit from the viewpoints and practical suggestions presented in the text. As one way of being selective and optimizing the investment of time, this reviewer suggests that the prospective reader cover the major part introductions and consult the readers' guide before turning to specific chapters.

JOHN L. WASIK  
*North Carolina State University  
at Raleigh*



H. J. Eysenck. *The Inequality of Man*. San Diego, California: EdITS Publishers, 1975. Pp. 288. \$8.95.

The thesis of this book written for the educated layman is that variability in human traits is primarily genetically determined. Eysenck is concerned with the origins of differences among people, regardless of race or ethnic background. (The terms "race" and "ethnic" do not even appear in the index.) The popular slant of the presentation is indicated by the size of the bibliography: only 79 references are cited.

As usual, Eysenck's arguments and evidence are conclusive; his position appears to be strongly supported while the environmentalists are generally discredited as persons overwhelmed by a desire for social reform and, thus, suffering from an inability to recognize the truth in scientific studies. Like a good lawyer, Eysenck presents the case for genetic inequality by emphasizing the favorable evidence and selecting for critical analysis those investigations most embarrassing to the opponent. His unrelenting attack includes unflattering characterizations of the rival, e.g., "extreme environmentalists," "a determined egalitarian," "convinced environmentalists," etc. And he is a master of the use of persuasive techniques; his most effective ploy is to occasionally concede a minor point to the opponent in order to give the impression that his presentation is truly unbiased.

On the other hand, Eysenck continues to be one of the most effective translators of the results and implications of scientific psychology into a form digestible by the interested layman. This recent volume may be viewed in the tradition of *Uses and Abuses*, *Fact and Fiction*, *Know Your Own IQ*, etc. Eysenck's explanations of heritability, test validation, research design, regression to the mean, and other technical topics communicate accurately to the non-specialist the meaning of what are generally considered to be fairly advanced issues in psychology. Finally, any tedium which remains is eased by Eysenck's colorful writing style and use of eye-catching examples interspersed throughout the book, e.g., his description of a study of the relationship between cancer of the womb and circumcision (pp. 27-28), a sample of items from the m/f scale of sex attitudes ("I would enjoy watching my usual sex partner having intercourse with someone else," p. 30), a correlation of  $-.63$  between IQ and number of teeth missing (p. 78), his amusing comparison of the consequences of a "mediocracy" with a meritocracy (p. 222), etc.

Approximately one-half of the book is concerned with intelligence (three chapters), with one chapter summarizing personality-related topics, and the final chapter reserved for a review of the social and political implications of the conclusions reached. Chapter one (Equality and Individuality) introduces the reader to Eysenck's point of view and includes brief synopses of three previously published books of similar title (authored by Rousseau, T. H. Huxley, and J. B. S. Hal-

dane); Eysenck thus puts himself in historical perspective! Chapter two (What Do IQ Tests Really Measure?) is a defense of psychometric assessment of intelligence—Spearman's *G* in particular. Eysenck has a predilection for drawing parallels between measurement problems in psychology and those previously encountered in the physical sciences, e.g., the measurement of intelligence is compared to the development of the thermometer; Spearman is compared with Dalton, the founder of modern atomic chemistry; the physical concept of "hardness" is compared to the psychological construct of intelligence, etc. Intelligence is demonstrated to correlate with "success in life" (income, occupational prestige), the evidence for creativity as an independent construct is found deficient, and the research on evoked potentials suggests a physiological basis for intelligence, all leading to the conclusion that the search for alternatives to *G* intelligence has failed.

The third chapter (Intelligence and Heredity) summarizes the evidence regarding the relative importance of hereditary and environmental influences in accounting for variation in intellectual functioning. The logic of twin studies and other methodologies are explained. Eysenck reviews studies of identical twins raised separately, fraternal twins, relatives, inbred individuals, orphanage children, adopted children, retardates, and environmental factors. The Milwaukee Project and Ellis Page's critique are summarized. The chapter concludes with consideration of two alternative explanations to the genetic hypothesis—malnutrition during infancy and social disadvantage: studies of Dutch children exposed to famine and "deprived" Eskimos are reviewed to rebut these proposed explanations. Chapter four (Intelligence and Social Class) continues the devastation of environmental explanations of individual variation, e.g., genetic forces guarantee a redistribution of intelligence (and, thus, opportunity) over generations, while environmental factors tend to perpetrate class distinctions; the relationship between intelligence and social class for fathers and children is consistent with the genetic hypothesis; education has little impact on pupil performance relative to intelligence ("Education is second only to psychoanalysis in making claims for untested and untried methods . . . (p. 151)"); etc.

Chapter five (Personality, Mental Illness, and Crime) presents an evaluation of the role of heredity in the causation of criminal behavior, mental illness, and variation in normal personality functioning. After reviewing nine twin studies, Eysenck reaches an uncharacteristically moderate conclusion: heredity and environment are about equally important determiners of criminal conduct. He even becomes temporarily humble: "I am not competent to argue . . . (p. 173)" and ". . . but this is a personal view and not relevant . . . (p. 173)." Reviews of twin and family studies of neurosis, alcoholism, and schizophrenia all provide support for a substantial hereditary component. A few studies of the P, E, and N scales of the EPI lead Eysenck to the conclusion that heredity ". . . accounts for not less than 50% of the total variance, and

may account for as much as 70% (p. 201)." Eysenck enlivens the chapter by attacking Laing's family oriented theory of schizophrenia and taking on the Women's Movement by arguing that psychological differences between the sexes are biologically based.

The last chapter (Social Consequences) considers various "implications of the facts surveyed in the previous chapters." The chapter begins with an analysis of Herrnstein's well-known syllogism and its corollaries: some are "undoubtedly true" while Eysenck's evaluation of others indicates that "Herrnstein is definitely beginning to run off the rails in his predictions (p. 217)." (Herrnstein may have underestimated the magnitude of regression effects.) Several examples are given which illustrate the potential dangers in disregarding the facts of biological inequality, e.g., the elimination of IQ tests in selecting children for higher education in Britain and the enforcement of affirmative action programs in the U. S. will result in lower quality of education for all. Eysenck's basic premise is that psychology can contribute to the improvement of the human condition only if inequality is recognized and differential treatments are viewed as fair for all; Jensen's level I and II abilities distinction and Sarason's modeling research with delinquents are used as illustrations. A potpourri of topics comprise the remainder of the chapter, e.g., diabetes, psychosis, and asthma are examples of genetically caused defects which are amenable to preventative treatments; Atkinson's CAI models illustrate the quantification necessary to objectify political decision-making; vocational satisfaction is viewed in terms of temperamental suitability for the job; etc.

At several points in the book Eysenck reassures the reader that he is not an extreme hereditarian and that he is only trying to establish a balance between the relative importance of hereditary and environmental causes of behavior, e.g., "We must learn to recognize the importance of interaction in regard to all aspects of behavior; it clearly will not do to slight either the importance of heredity or that of environment (p. 242)." On two separate occasions he makes the critical point that heritability estimates currently available are based on naturally occurring variation in the environment and, thus, cannot be generalized to behavioral treatments or programs which extend beyond the normal limits.

Nevertheless, since the book was aimed at the layman, who probably is not familiar with Eysenck's inclination to engage in controversy and his talent for presenting an impressive case for his point of view, a final chapter or rejoinder written by "a determined egalitarian" would have done much to promote the balance that Eysenck hopes to achieve.

BRIAN BOLTON  
*University of Arkansas*

Gene V Glass, Victor L. Willson, and John M. Gottman. *Design and Analysis of Time-Series Experiments*. Boulder, Colorado: Colorado Associated University Press, 1975. Pp. xi + 241. \$10

For a good many years the field of design and analysis of behavioral science experiments has changed primarily by accretion. That is, there has been more and more development of what is currently referred to as the "randomized comparative experiment." One of the first breaks with what has become a major tradition was Campbell and Stanley's *Experimental and Quasi-Experimental Designs for Research* (1966), which examined several designs outside of the mainstream framework, in particular, the "interrupted time-series experiment." Research utilizing these procedures appeared slowly through the 1960's, but of late has begun to surface more and more frequently. Glass, Willson and Gottman's book deals solely with time-series experiments and marks a major change in our current experimental design and statistics tradition. It will, without doubt, be looked back upon in future years as a watershed publication. Prophecy is always dangerous, but several factors make it a safe bet that within a decade one-third to one-half of the material currently taught in graduate experimental design and statistics courses to students in the behavioral sciences will have been replaced by material similar to that in this book. Some of the reasons for this change are obvious, others are more subtle.

In the first place, the very extent of the development of randomized comparative experiments has focused attention on the fact that they are best at appraising the results of treatments inserted into an experimental medium at a particular point in time and manifesting results immediately thereafter. But, in social systems as well as in the lives of individuals, one is more likely to see patterns and gradations of experimental effects rather than simple, sharply-delimited expressions of them. Time-series analysis allows a wide-angle view of experimental effects. Second, single subject designs in the operant conditioning field or in any of a number of areas utilizing behavior modification techniques have been largely without an explicit design and analytic framework. Time-series analysis provides one. Third, single subject designs or single group designs are becoming vastly more popular in all areas of psychology and in most applied work in education because of increasingly greater difficulties in obtaining sizeable, multiple groups of subjects and because of the growing unwillingness of school administrators to go along with randomly equating classroom groups to be evaluated for the effect of different treatments. Time-series analysis again provides a convenient framework for the analysis of such studies. Finally, a brief look at any of the dozen most popular "statistics and design" books shows that almost none of them contains an appreciable amount of material related to basic design considerations. They are, to be sure, top heavy with the type of rationale that justifies factorial arrangements of treatments, but the basic ideas of experimen-



tal design receive almost no attention. This frustrates teachers and students alike. Time-series analysis goes off on such a disparate tangent to the traditional approaches that a great deal of time must be devoted to design considerations. This has great appeal to graduate students and makes the statistical and analytical procedures that are included seem much more appealing. For all of these reasons, the field of time-series experiments is one of the waves of the future.

A time-series is basically a set of observations on some dependent variable taken in sequence, each measurement separated from the other by a (generally) fixed amount of time. In the simplest time-series experiment, one repeatedly observes a subject or a group with respect to some dependent variable. After a number of observations, a treatment of some sort is imposed, and observations are then made through a number of subsequent periods to determine whether the level of the dependent variable remains the same or changes in some systematic fashion. The major analytic problem is that there are, of course, fluctuations in the level of the dependent variable both before and after the intervention; and one has to determine whether changes subsequent to the intervention are different from what could be expected on the basis of the random shifts inherent in the series.

The primary design considerations in time-series experiments are the same as in any research problem, but there are several ideas which receive particular emphasis. They revolve around whether the same group or different groups are observed at different points in time, whether more than one group is observed simultaneously, whether one intervention or more than one is used with each group, and whether the effects of the two or more interventions have been observed in reverse order to enable assessment of interaction or sequence effects. Interaction effects are viewed from a different perspective than in randomized comparative experiments: the effects on the dependent variable become more complex since one does not examine them at a single point in time but at many points in time. Rate of change, magnitude of change, or even lack of change will each be interpreted differently depending upon the timing of the intervention, the duration of the intervention and temporary occurrences in the experimental field. Glass et al. present an extremely thorough and, on the whole, quite readable discussion and evaluation of various time-series designs and interpretation problems. The level of discussion is kept admirably even, and should be intelligible to any interested student. A great many examples from the literature of a variety of fields are examined in the context of each design question. Another strength of the discussion is that it is critical: the multiple baseline design introduced by Risley and Baer in 1969 which is so widely used in the field of operant conditioning is shown to depend upon a logic that may be asymmetrical or even contradictory. The authors are not shy about discussing and naming the many factors which may work to invalidate time-series experiments. They discuss in detail such factors as historical invalidity,



reactive intervention, multiple intervention interference, instrumentation problems and the like.

The mathematical approach utilized by Glass et al. derives from a development of spectral analysis due primarily to Box and Jenkins (1970), and involves linear models collectively termed "auto-regressive integrated moving average models," or "ARIMAs." The simplest cases are an auto-regressive model and a moving averages model. In simple terms, an auto-regressive model is one in which succeeding time states depend upon preceding ones. Actually, of course, the dependency may be over a time difference of several observational periods. If the relationship is from one observational period to the next, this is called a lag one process, if from the first observational period to the third, it is called a lag two process, and so forth. The moving averages model essentially assumes that the state of a process at a given time is the average of a series of random shocks delivered at preceding periods. Thus, while there is no significant correlation between one time period and the next, the present state of the series is an average of what has gone before.

A problem in analyzing time-series is the fact that some are "non-stationary," that is, they tend to wander temporarily from one overall level to another and back again in stochastic fashion. Taking differences between successive time points (lagged by one, two, three or more time periods) will generally serve to convert a non-stationary process to a stationary one which may then be described in terms of an auto-regressive or moving averages model (or a combination of the two).

The ARIMA schema introduced by Box and Jenkins and capitalized upon here by Glass et al. has the conceptual simplicity and beauty of allowing the description of a time-series simultaneously in terms of whether it is auto-regressive, requires differencing, and/or has moving averages components. The authors introduce both a notation and a calculus of operators for classifying and working with ARIMA models. The procedures are simple and straightforward for anyone with prior experience with linear models. The computations involved are extremely well set forth and well illustrated, and utilize both good real life examples and some excellent simulated data sets.

From the point of view of the experimentalist, the real problem in analyzing a time series is that of determining whether or not an intervention produces an effect over and above what might be accounted for by chance, and, if it does, estimating its magnitude. The procedures illustrated by Glass et al. require several steps. First of all, one identifies the model (both pre- and post-intervention). Parameters are estimated and the level of the series determined. The intervention effect is then estimated through a maximum likelihood procedure, which depends primarily upon minimizing the error sum of squares. This involves repetitive solutions of a set of linear equations by varying values of (ordinarily) two parameters until a minimum is found. Again, the procedure and computations are well illustrated. Moreover, the au-

thors have developed a set of computer programs to facilitate computations which should bring use of the model quickly within grasp of many prospective users.

The next to the last chapter of the book deals with analysis of related time-series through concomitant variation. That is, if one series is related to another, one may often discover useful information about one from known characteristics of the other. This procedure is in many ways very similar to co-variance analysis and will thus have a familiar ring for many behavioral and social scientists.

The last chapter of the book deals with several specialized topics including deterministic drift, changes in variance, changes in the model, and cyclic variation. Although they are short, these sections are well written and useful.

There are appendices which deal with spectral analysis of time-series and with the linear model and least squares theory. Both of these appendices are too compressed to be of use to anyone who does not have the fundamentals of those techniques reasonably well in mind to begin with. Another appendix contains listings of the data sets used to illustrate particular problems and should, together with the programs which the authors have written, be most useful.

The formal aspects of the book are rather unexceptional except that the manuscript has been typed and photographed so that the right margin is not justified. A good many of the "instant books" produced in this fashion are full of typos and are poorly edited, but the Glass et al. volume does not belong in this category. It is well constructed and put together, there are few typos, the only bad one being in the third paragraph on page 48 where "unobserved" is substituted for "observed." One would expect the symbols and formulae to be exceptionally poorly set and hard to read, but they come through well. As stated above, the examples are excellent, and they are completely and accurately cited in an excellent reference section.

Glass et al. set out to show where new ground has been broken and to direct others not only in where to plow their own furrows but in how to keep them straight. On the whole, they have done an excellent job, and it is likely that this book will come to be viewed as a true landmark. It differs from many landmarks in that, although a first effort, it is exceedingly well organized and well constructed and should stand the test of time exceedingly well.

## REFERENCES

- Box, G. E. P. and Jenkins, G. M. *Time-series analysis: Forecasting and control*. San Francisco: Holden Day, 1970.
- Campbell, D. T. and Stanley, J. C. *Experimental and quasi-experimental designs in research*. Chicago: Rand McNally and Co., 1966.

JAMES A. WALSH  
University of Montana

John L. Hayman, Jr. and Rodney N. Napier. *Evaluation in the Schools: A Human Process for Renewal*. Monterey, Calif.: Brooks/Cole, 1975. Pp. xi + 143. \$3.95 (paperback).

This short, concise book may be read with profit by anyone concerned with evaluation in the schools. It is not a comprehensive treatise on evaluation, since many of the technical details about research design, statistical analysis, and instruments employed in evaluation studies are not presented. Matters such as how to survey, interview, select tests, and the like are not discussed at all. Rather, what the authors stress are procedural matters and the human factors involved in evaluation, including process and subjective evaluation. Also emphasized are the facts that evaluation can be helpful or harmful, and it is the responsibility of the evaluator to determine which it shall be.

The authors possess a penchant for outlining and schematizing, a talent that will facilitate the use of the book as a sourcebook and a text but which results in somewhat slow reading. As stated in the Preface, this is not a cookbook; it is an attempt to prepare the would-be evaluator through means of practical examples. Such examples are less plentiful in the first four chapters, which are more like expanded outlines of principles and procedures, than in the last four. Furthermore, the authors maintain that the book is appropriate for undergraduates in educational psychology and related courses, as well as educational practitioners. Certainly many undergraduates in educational fields could study the book with benefit, but most will probably find it fairly difficult reading. This reviewer considers the book's primary value to be in training educational evaluators—when supplemented extensively by books on educational measurement, administration, and psychology.

The eight chapters comprising the book are entitled:

1. The New View: Evaluation as Integral to the Educational Process
2. Planning for Evaluation
3. Goals and Objectives in School Evaluation
4. Evaluation in Program Development
5. Process Evaluation: Program Development and Organizational Applications
6. Implications of Process Evaluation for the Classroom Teacher
7. Accountability and Evaluation in the Schools
8. Synopsis: A Brief Guide to the Evaluation Process

Every chapter is well written, if succinct, and contains a concluding summary.

Throughout the chapters the point is repeated that feedback of the results of evaluation from evaluators to users is essential if evaluation is to be more than a futile exercise in wasting time. The first four chapters are highly structured, describing methods of planning, effecting, and utilizing evaluation, as well as input, process, and outcome

variables. Beginning with Chapter 5 the orientation of the book becomes more psychological, or rather socio-psychological. The descriptions of group dynamics and process evaluation in Chapters 5 and 6 are especially good. Chapter 7 on accountability deals more with social psychology and management than with technical issues concerning the measurement of change. There is a good discussion of Dyer's multiple regression approach to accountability and a recognition of its shortcomings. The weaknesses of the multiple regression approach, and the necessity of complementing it with a "management-by-objectives" approach, are discussed. The latter approach, of course, also has weaknesses: getting participants to state objectives clearly; extensive administrative time required; the fact that the approach sometimes results in psycho therapeutic soul-searching more than realistic attempts to state and attain objectives.

In Chapter 8, a comprehensive summary of the previous seven chapters, the authors continue to make important points concerning such matters as the role of environment in evaluating affective objectives; the value of judgmental evaluation information; the use of teacher-researchers in planning, effecting, and evaluating the results of research in the schools; and the necessity of creating understanding and rapport on the part of the teachers, administrators, and others affected by evaluation.

LEWIS R. AIKEN, JR.

John C. Loehlin, Gardner Lindzey, and J. N. Spuhler. *Race Differences in Intelligence*. San Francisco: Freeman, 1975. Pp. xii + 380. \$12.00 and \$5.95 (paperback).

This superb volume should be the final word on origins of race differences in intelligence. Not because it provides any ultimate answers, but for exactly the opposite reason. All evidence currently available has not really brought us much closer to any definitive conclusions. In fact, this comprehensive statement of our ignorance will probably serve to trigger an avalanche of investigations. The authors virtually predestine this event by outlining 10 "promising areas of research" in the last chapter.

*Race Differences* was prepared under the auspices of the SSRC's Committee on Biological Bases of Social Behavior. The authors completed their first year's work on the book in the intellectually stimulating environment of the Center for Advanced Study in the Behavioral Sciences. They were assisted by an advisory board consisting of 20 distinguished scientists, educators, and public figures (from Anastasi to Wolffe). A draft of the manuscript was reviewed by advisory board members, by six minority-group consultants, and—in various degrees of thoroughness—by 50 prominent academicians in the biological and social sciences (including Eysenck, Jensen, and Shockley).



The authors' stated goal was "to provide a sober, balanced, and scholarly examination of the evidence" regarding the relative contributions of genetic and environmental variations in explaining racial-ethnic IQ differences and to discuss the social and political implications of the results. Their purpose is impressively achieved in this *tour de force* of scholarship. The book is extremely well organized and beautifully written. Every conceivable source of relevant data is reviewed. The evaluations of the investigations are insightful and thorough. Summaries and discussions are strategically placed in the text to have maximum relevance and impact. Technical issues are handled in 70 pages of appendices which are marvels of clarity and comprehensiveness.

And while the authors' conclusions are almost always multiply-qualified and hedged, they point out repeatedly that the inconsistency and poor quality of the evidence necessitates their extreme tentativeness, e.g., regarding nutrition and black-white IQ differences: "The best evidence needed to answer these questions is lacking, and the relevant, usable information is scanty, and of unknown or dubious reliability (p. 225)."

*Race Differences* consists of 10 chapters which are organized into 3 major sections: Issues and Concepts (four chapters), The Empirical Evidence (four chapters), and Conclusions and Implications.

Chapter one (The Problem and Its Context) traces the roots of "the controversy" to Darwin (evolution), Galton (inheritance), and Mendel (genetics). Intelligence tests are characterized as "one of the most significant technological accomplishments of the social sciences (p. 5)." The key role of Jensen's 1969 *HER* article, the involvement of subsequent protagonists (Herrnstein, Eysenck, and Shockley), and the Coleman report are reviewed. The chapter concludes with eleven capsule summaries of "persistent misconceptions" which serve to introduce terminology and preview later chapters.

Chapter two (Race as a Biological Concept) presents a brief overview of the concept of race, illustrates racial variation in color and size using North American house sparrows, and reviews race formation in prehistorical man. The authors rephrase the central question of the book—Are the genes that determine intellectual capability differentially distributed among the major races?—and evaluate the four principal evolutionary mechanisms (mutation, drift, migration, and natural selection) which might account for differences in gene-frequencies between racial groups. They conclude that differential natural selection "could result in substantial differences between large human populations in the distribution of the genes underlying general intelligence (p. 48)." A hypothetical numerical example is used to demonstrate the potential effects of natural selection.

In chapter three (Intelligence and Its Measurement) three broad conclusions about the construct of intelligence are outlined: (1) it is a measure of performance relative to a particular standardization popu-



lation, (2) the choice between G and a multi-ability conceptualization is a matter of convenience and purpose, and (3) it is not innate, but, rather it is developed. Reasoning by analogy using the trait of stature leads to the suggestion that intelligence may have been differentially selected in different environments in the past. Finally, three studies of test-bias reversal (rural children, Indian children, and blacks) suggest that IQ tests may be biased against some groups. Not all studies support this conclusion, however.

Chapter four (Heritability) is a rather technical introduction to a number of complex topics, including developmental genetics, covariance and interaction, genetic variance components, distinctions among heritability coefficients, comparisons of models of inheritance, heterosis, etc. Among the major conclusions are the following: (1) Within-population estimates of IQ heritability *may*, but *need not* have value for interpreting between-population (racial or cultural groups) differences in IQ-test performance. This point was the basis of Kagan's criticism of Jensen's *HER* conclusion concerning possible racial differences. (2) The broad heritability of IQ in Caucasian populations lies in the range from .60 to .80. Two prominent exceptions to this conclusion are the estimates calculated by Jencks and Kamin. While Jenck's estimate of .45 poses no serious problem, Kamin's assertion of zero heritability, if supportable, would severely damage the authors' argument. Therefore, they thoroughly review Kamin's evidence, analyses, and arguments and convincingly demonstrate that his conclusions are based on a biased, selective review that suffers from several logical and statistical errors.

Chapter five (Genetic Designs) presents a detailed review and evaluation of (1) twin and sibling studies which provide comparable heritability estimates for black *and* white samples, (2) interracial adoption studies, (3) studies of half-siblings, and (4) studies of subjects of mixed racial backgrounds, e.g., correlation of skin color and intelligence, correlation of blood group genes and IQ in blacks, offspring of black soldiers in Germany in WWII, etc. Although every one of the dozen studies reviewed suffers from one or more defects—sampling limitations, incomplete information, and various other methodological flaws—the authors conclude that (1) IQ is substantially heritable in the US black population, and (2) the existing genetic evidence could be used to support either environmentalist or hereditarian viewpoints.

Chapter six (Temporal Changes in IQ) consists of a review and evaluation of studies of population trends in IQ, developmental studies of IQ change, and the effects of compensatory education programs. The authors draw on a wide variety of data sources, e.g., Binet and WISC standardization data, the Coleman report, US Army induction test data, etc., to reach three conclusions: (1) population increases in IQ are associated with educational improvement, (2) black-white IQ differences emerge by age three or four and generally remain fairly stable throughout the school years, and *may* increase during adult-

hood, and (3) the massive research effort in compensatory education has not produced any breakthroughs. The authors summarize by stating that while environmental factors do influence IQ development, this does not rule out substantial genetic determination of individual or group differences.

Chapter six is notable for two reasons. It is the first to include data (several tables) which document the black-white IQ difference of one standard deviation referred to initially in chapter two. The fact that there is no chapter devoted to the establishment of the accuracy and meaning of this performance difference suggests that the authors have considerable confidence in the available evidence and measuring instruments (but see their major conclusion in chapter ten!). The second point concerns the generally positive evaluation of Jensen's position. In fact, some will view *Race Differences* as a vindication of Jensen. On several occasions the authors imply that Jensen has been misread or misunderstood. In an interesting (if not intentional) phrasing, they conclude that the failure of compensatory education to ameliorate black children's scholastic deficits "*made it not unreasonable* for Jensen (1969) and others to reopen the question of a possible genetic component (p. 162)." The italics in the quote are added to emphasize the similarity to Jensen's original statement (1969, p. 82). Finally, Jensen is cited 22 times in the text (Burt, Nichols, and Shuey have nine citations each) and has 16 entries in the references (a neurochemist, Dobbing, is second with seven entries).

Chapter seven (Cross-group Comparisons of Intellectual Abilities) continues the review of studies which bear on the main question addressed in the book. The most consistent evidence so far concerns the structure of abilities in various racial-ethnic groups: the underlying dimensions are the same. The investigations of ability profiles of various racial-ethnic groups (which might be hypothesized to be the effect of differential selection and adaption to different environments) are much more ambiguous. In general, there is evidence of differences in levels of performance and that blacks do better on verbal than non-verbal tasks—but there is little evidence for a specific perceptual deficit in blacks. The chapter concludes with an excellent discussion of the often misunderstood "regression" phenomenon and a good summary of Jensen's theory of abilities.

Chapter eight (Nutrition and Intellectual Performance) begins with an introduction to the principles of nutritional science. The evidence suggests that short periods of malnutrition early in life may result in permanent and irreversible intellectual deficit. Furthermore, prenatal nutrition, birth weight, and IQ appear to be related in a causal fashion. And while blacks and other U.S. minorities are less well nourished on the average than whites, the authors calculate that the cumulative effect may account for "a few points" of the black-white IQ difference.

Chapter nine (Summary of the Empirical Findings) simply lists 20 "empirical generalizations" which the authors derived from evidence

reviewed in chapters five through eight. The distribution by topics is: quality of evidence (2), heritability of IQ (3), racial mixture (5), comparisons of socioeconomic and racial groups (7), and nutrition (3).

Chapter ten (Implications and Conclusions) begins with a statement of the authors' conclusion regarding the sources of observed IQ differences among racial-ethnic groups: the differences are due to (1) psychometric deficiencies in test instruments, (2) differences in environmental conditions, and (3) genetic differences. The relative weight assigned to each component is simply a matter of judgment! The authors are fully aware of the indeterminacy of their major conclusion and explain that the evidence does not support any stronger statement. A few pages later they do make a slightly stronger assertion: "We consider it quite likely that *some* genes affecting *some* aspects of intellectual performance differ appreciably in frequency between U.S. racial-ethnic groups . . . (p. 240)."

So what do we really know about the origins of racial differences in IQ? Surely, there is something wrong when the strongest conclusion that is scientifically warranted is virtually meaningless. The authors have convinced me that the question is simply unanswerable at the present time. Should research on racial differences in intelligence continue? There are too many socially important topics to be researched—and no conceivable conclusion concerning the source of black-white IQ differences is going to alter the fact that it is the *individual*, not any particular population subgroup, about whom decisions are made and for whom opportunity exists in a democratic society. The authors continually stress this point and the well known fact that much more IQ variation occurs within racial-ethnic groups than between them; both points would seemingly detract from the importance of research addressing the primary question of the book.

The remainder of chapter ten consists of two excellent sections: (1) a discussion of the social and political implications of *possible* genetic differences in intelligence, and (2) an assessment of the social context of "sensitive" research, followed by brief descriptions of ten *promising* areas of research on racial-ethnic differences. The authors affirm their belief in the importance of basic research in general, and judge (some) research on racial differences to be of fairly high scientific and social priority. Furthermore, they support Shockley's (or anyone else's) right to investigate possible dysgenic trends within racial groups, but accord it low priority.

As a reviewer of *Race Differences* I feel an obligation to emphasize the distinction between my evaluation of the authors' effort and product—probably the most scholarly volume I have ever read—and my reaction to their conclusions—disappointment. Of course, they cannot be held responsible for the inconsistency and poor quality of the evidence. Nor should the many investigators whose research was reviewed be blamed—it is an extremely difficult question to study. My personal feelings about research on racial differences are perfectly

stated by Loehlin, Lindzey, and Spuhler: "When one considers the hundreds of inconclusive comparisons of black and white IQs that have appeared in published studies in the scientific literature in the last several decades, the thousands of pages of print devoted to discussion of race and IQ, and the massive commitments of human and financial resources based on *assumptions* about matters to which these studies are directed, one can hardly help but feel some serious misgivings about the responsiveness and responsibility of the system of rewards and support that determines which research gets done in the social sciences (p. 256)." Amen.

BRIAN BOLTON  
University of Arkansas

Elijah P. Lovejoy *Statistics for Math Haters*. New York: Harper and Row, 1975. Pp. x + 251. \$8.95 paperback.

One might paraphrase the author to indicate more fully his intent and approach as "The Statistics of Psychological Experimentation for Undergraduates Initially Aversive to Mathematics." This would be true as indicated later in this review, but it should not be taken to imply limitation of the use of this publication to individuals in such circumstances.

This textbook/manual benefits from its explicit focus on those being introduced to the tools of statistical logic in an undergraduate major in general psychology. We may accept the author's and editor's assurances that this material has met the challenge of field testing. It presents the logic of inferential statistics well by reinforcing at successively more advanced levels the initial presentation built around the signs test and coin-tossing. In this respect, it is reminiscent of Walker and Lev's *Statistical Inference*, which I still find useful in teaching a second course (in inferential statistics) to graduate students in education. In the light of that experience, this approach may be judged quite as helpful to potential psychological researchers with strong mathematical backgrounds, because I have frequently found these individuals just as hard put to deal with probabilities, as distinguished from exact mathematics, as are their mathematically less proficient classmates. The concepts of Type I and Type II errors receive balanced treatment early. The emphasis on understanding statistics as a means of analyzing and evaluating data to reach decisions, rather than as an assemblage of techniques leading to solution of numerical problems, is wholesome and warrants a high rating of the book as a textbook for psychology majors beginning the study of statistics.

The remainder of this review must be devoted to indicating the usefulness of this book in the diverse courses in statistics offered in different fields of applied social science: education, economics, sociol-



ogy, etc. Here it would have supplementary use in the initial stages of any course in inferential statistics. Those who prefer to teach descriptive statistics first, as a course worthwhile in itself for teaching means of summarizing and presenting quantitative information in readily understandable form, may well continue to reserve inferential statistics for a second course. Taking that approach and using this publication as a supplementary workbook in the second course should provide the added value of an independent source of stimulation about the thinking processes involved. In these applied fields many of the studies fall into categories variously called quasi-experimental, *ex post facto*, or causal-comparative, rather than rigorously experimental, but the logic of experimentation is basically the same and the illustrative examples would be clarifying.

The sequence of topics in Part III, following the earlier treatment of the signs test, binomial and normal distributions, does unnecessary violence to systematic relations, by going from the *t*-test for single means and correlated means, into correlation, chi-square, and then the *t*-test for means of uncorrelated samples, concluding with the difference between two-tailed and one-tailed tests. Putting aside the advantage a separate course in descriptive statistics provides for organizing concepts and measures under the percentile and moment systems, a structure for paralleling treatment of discrete and continuous variables, respectively, gives a meaningful relation among approaches in inferential statistics that is lost by purely topical treatment. And even "math haters" gain something by such structuring.

It should be noted that analysis of variance (and covariance) is omitted as beyond the scope of the beginning course. This is regrettable because the logic of experimentation which dictates that ANOVA must precede study of mean differences when several means are available, is lost. Of course, in a statistics sequence for psychology students who had used this book, these topics and the logic of multivariate analysis, discriminant analysis, factorial designs, etc. would be covered in subsequent courses.

The use of flash cards to acquire concepts and formal definitions should help beginning students, as should also the inclusion of partial statistical tables in the body of the text. Perhaps they are not essential, but they are probably helpful. (By way of contrast the minor exercise on the relative merits of \$8 and \$5 Scotch is dubious on about every count.)

In sum, the book seems well designed for its avowed purposes and to have additional value for possible supplementary uses.

WARREN G. FINDLEY  
*University of*  
*Alabama in Birmingham*



Jum C. Nunnally. *Introduction to Statistics for Psychology and Education*. New York: McGraw-Hill, 1975. Pp. x + 342. \$10.95.

There is a veritable glut of textbooks in certain areas of psychology and education today. General psychology and educational psychology are two such areas, and introductory statistics is fast becoming another. Given this state of affairs, it behooves an aspiring author of a textbook in any of these areas—whether his interest is pedagogy, profit, or both—to look carefully at what is already available and find a way of doing it better, if he can. This is what the author of the present introductory statistics book has tried to do. He apparently feels that there are enough statistics cookbooks, mathematically-oriented statistics books, and books concerned primarily with statistical inference rather than description. Verification of this supposition is not difficult to find; one need search no further than McGraw-Hill's list of a dozen statistics books.

In a way, it is good to have textbooks of varying genre available; it gives instructors a choice. Unfortunately, instructors often seem to select textbooks for themselves rather than their students. And the fact that an instructor likes a book is obviously no guarantee that his students will follow suit. Instructors have even been known to choose two textbooks for a course, taking their lecture notes from the better of the two and assigning the remaining book as the course text!

The yellow-covered volume that is the subject of this review could presumably serve either of these functions. It consists of 13 chapters grouped into four parts: Fundamental Concepts, Descriptive Statistics, Inferential Statistics, and Nonparametric Statistics. I didn't quite understand the purpose of the illustration on the front cover—a profile of a human head with an irregular line graph running sagittally through it, but that is obviously less important than understanding what follows the cover. What does follow is a highly verbal book containing no statistical symbols before Chapter 5. At that point, however, symbols begin to fall thick and fast and the coverage becomes quite condensed. Measures of central tendency and variation are completed in one chapter, followed in rapid-fire succession by norms, correlation, *t* tests, *F* tests, analysis of variance, and a bit about nonparametrics. Principles, rather than mathematics, constitute the forte of the book. But occasionally the author betrays his area of specialization, for example when he derives formulas pertaining to the correlation coefficient and takes many of his examples from the field of psychological testing. This is to be expected, since it is difficult not to dwell on what one knows best. For this reason, statistics books written by experimental psychologists tend to stress inference and experimental design. In contrast, those written by specialists in psychological and educational measurement, such as Nunnally and Guilford (see Guilford & Fruchter, 1973) give more attention to correlational analysis and norms.

The author of the present textbook had three collaborators—Robert L. Durham, L. Charles Lemond, and William H. Wilson. Nevertheless, several typos, errors and inconsistencies sneaked through. For example, formula 5-6 on p. 104 is incomplete, and the definition of *percentile* is patently wrong, viz. "the percentage of persons who fall below a particular score" (p. 119). The correct definition of the  $p$ th percentile is "... that value on the scale of measurement below which  $p$  percent of the cases in the distribution fall, ... whereas its corresponding percentage is known as the *percentile rank*." (McCall, 1975, p. 64) An inconsistency occurs when, in the preface, the author disparages the inclusion of "museum pieces" in statistics texts, and then proceeds in Chapter 5 to describe an archaic statistic known as the average deviation. Another inconsistency occurs in defining  $s^2$  in two ways—with  $N$  and  $N-1$  in the denominator. This double usage will probably create confusion for the introductory student who reaches the section in Chapter 9 on computing the standard error of the mean.

In the preface the author stated that only topics that are widely encountered in psychology and education are discussed in the book. To be certain of this, one would need to tabulate the usage of certain statistical methods in articles appearing in professional journals in these fields. If this were done, I seriously doubt whether the average deviation, or even the mode, would qualify for inclusion in the class of most frequently used statistics. On the other hand, the binomial probability formula, the test for difference between proportions, partial correlation, test of significance of regression coefficients, the test that two correlation coefficients are equal, and even analysis of variance by ranks are methods used in many published investigations but omitted in the present textbook. Finally, although this book purports to emphasize the understanding of principles and concepts rather than mathematical skills, I missed any reference to the important topics of the central limit theorem, Type I and Type II errors, the power of a statistical test, the consistency and bias of estimators, or even the descriptive statistic of kurtosis.

In spite of these shortcomings, it can be concluded that, to a certain extent, the author and his collaborators have achieved their objectives—to write an interesting book emphasizing understanding rather than skill, including methods used most widely in psychology and education, dealing with description more than inference, and trying not to overwhelm the student with mathematics and computation. Whether or not these objectives are the best ones to follow in writing an introductory statistics text is debatable. For example, interest and understanding are often obtained at the expense of comprehensive survey and in-depth coverage. It can also be argued that since the first course in statistics is the only one that most students will take and even those who go on to advanced courses benefit from a thorough survey in the first course, the beginning statistics textbook should be comprehensive. It should not only give a thorough coverage of fundamen-

tals but also serve as a sourcebook and reference manual long after 90% of the course experiences have been relegated to the same netherworld as other college educational experiences.

Authors of statistics textbooks sometimes try too hard to minimize "symbol shock" in the mathophobes who are forced to take statistics in order to be certified to "work with people." Although social science and humanities majors frequently suffer from an inability to manipulate numbers and algebraic symbols, a more serious deficit may exist in their ability to do precise abstract thinking of the problem-solving type. This is especially evident when students do fairly well with the descriptive aspects of statistics but begin to have headaches when the topic of statistical inference is introduced. It is difficult to prevent or cure such headaches; even when a predominantly verbal text is used, the concepts still require logical, reflective thinking.

### REFERENCES

- Guilford, J. P. and Fruchter, B. *Fundamental statistics in psychology and education*. (5th ed.) New York: McGraw-Hill, 1973.
- McCall, R. B. *Fundamental statistics for psychology*. (2nd ed.) New York: Harcourt Brace Jovanovich, 1975.

LEWIS R. AIKEN, JR.

William H. Sewall, Robert M. Hauser, and others. *Education, Occupation, and Earnings, Achievement in the Early Career*. New York: Academic Press (A Subsidiary of Harcourt Brace Jovanovich, Publishers) 1975. Pp. xviii + 237. \$16.50.

This important longitudinal research reports the comparative cause-and-effect relationships between occupational achievement and earnings as effects during the first 10 years after graduation of a large sample of young men who graduated from Wisconsin high schools in 1957. The causal factors included father's educational attainment, mother's educational attainment, status of father's occupation when son graduated, parents' average income 1957-1960, and son's Henmon-Nelson measure of mental ability.

The effects include this measure of mental ability, the son's educational attainment, the son's occupation, the son's canonically weighted average of 1965-1967 earnings. The father's and son's occupations are in terms of Duncan's socioeconomic index, (SEI).

In 1957, 36,171 persons graduated from parochial and public high schools in Wisconsin and 94.4% responded to a questionnaire prepared by Professor Kenneth Little of the University of Wisconsin with the cooperation of the State Superintendent of Schools. From the set of 34,151 questionnaires, a random sample of 10,750 males was se-

lected later reduced to a "cohort" of 10,317 cases for which Henmon-Nelson measures of mental ability were available.

Path coefficient or path analysis diagrams were drawn to illustrate the relationships between the variables listed above. The numerical entries in these diagrams are path coefficients or regression coefficients in standard score form. The squares of these coefficients are coefficients of determination. These may be used to indicate the proportions of an effect variable to be attributed to various causal ones.

Sewall and Hauser credit the path analysis procedure to Otis Dudley Duncan without mentioning its history which goes back to its origination by Sewall Wright in 1921 and its use since that date by a number of persons including this reviewer.

Sewall, Hauser, and their associates deserve commendation in studying the possible effects of nonresponse in biasing the data. They demonstrate that nonresponse had little effect on the general patterns of the 1957 and 1964 data. "Bias due to nonresponse has been shown not to affect materially either the univariate or the multivariate statistics of key variables." (p. 42).

Chapter 7 summarizes the relationships between socioeconomic background and mental ability as causes and educational, occupational, and financial achievements as effects. There is excellent critical discussion of social psychological factors, college effects, and ability-schooling interactions.

MAX D. ENGELHART

David M. Shoemaker. *Principles and Procedures of Multiple Matrix Sampling*. Cambridge, Massachusetts: Ballinger Publishing Company, 1973. Pp. xviii + 305. \$12.50.

This text serves as a reference manual for educational and psychological researchers and evaluators interested in the theory, development, and implementation of multiple matrix sampling designs. As such, it represents an organization and synthesis of the multiple matrix sampling literature through the end of 1973. According to the author, "Throughout [the] book an attempt has been made to keep the practitioner clearly in mind. The emphasis is clearly on the why, when, and how to use multiple matrix sampling." With its emphasis on practical techniques, the book is designed to make multiple matrix sampling more readily available to those interested in the assessment of group performance in a wide variety of areas.

The book is divided into two parts: In the first, the author introduces the multiple matrix sampling model, presents relevant theory, discusses applications, and considers guidelines for utilization. Specifically, the first part of the book consists of seven chapters spanning 85 pages of text and covering the following areas: 1. Definition, advan-



tages, limitations, and applications of multiple matrix sampling, 2. guidelines for the utilization of multiple matrix sampling, 3. computational formulas, 4. use of computer simulation techniques in designing multiple matrix sampling studies, 5. hypothesis testing, 6. applications, and 7. future possibilities.

While we are generally very positive about the coverage of topics in the first part of the book, we did uncover a number of weaknesses that we feel reduce the usefulness of the book. First, because the field of multiple matrix sampling is expanding at a tremendous rate, the book will be quickly dated. For example, important developments such as the use of multiple matrix sampling scores for examinee ability estimation (Bunda, 1973) and research on the problem of context effects (Feldt and Forsyth, 1974) were not available to the author. (In fact, recent research results on context effect are opposite to the results reported in the book and the tentative conclusions drawn by the author.) Thus, while the book serves as an excellent starting point for the study of multiple matrix sampling, an individual desiring an up-to-date comprehensive coverage of the field would now need to include new contributions to the field (e.g., Sirotnik, 1974) in his/her readings.

Our second and most serious criticism concerns the chapter on guidelines. Since the book is primarily intended for practitioners, the guidelines should be clear and detailed; instead the guidelines are presented in a general form. Also, little in the way of a rationale for the guidelines is presented and there is no mention of practical procedural guidelines such as those concerning economic constraints, and strategies for handling multiple item types and directions. One or more carefully worked examples showing the steps in designing and implementing a multiple matrix sampling study would have been informative.

The second part of the book includes chapter references, a very helpful bibliography on multiple matrix sampling, and a 217 page section describing two computer programs designed for use in implementing multiple matrix sampling studies. The first program has been designed to allow the user to estimate parameters in the multiple matrix sampling model. The second allows the user to simulate various multiple matrix sampling designs in different testing situations to study their effectiveness. The descriptions of the programs include summaries, lists of subroutines and functions, data specifications, program limitations, operating instructions, program flow charts, complete listings, and sample output. While the summaries, flowcharts, and sample output are informative, the usefulness of including 150 pages of computer program listings seems dubious. Computer programs are seldom obtained from a listing anyway, and when we consider this fact along with the knowledge that the inclusion of program listings added substantially to the cost of the book, we feel that the author's decision to include the listings was a poor one.

Overall we found the book to be well-written, appropriately organ-



ized, technically correct, carefully edited, and highly suited for self study. When the reader recovers from the shock of discovering that more than half the book consists of computer program listings, we think he/she will find it to be an important and useful contribution to the psychometric research field. Also, we feel sure that the book will achieve its fundamental purpose of furthering the appropriate implementation of multiple matrix sampling designs in a wide variety of areas.

## REFERENCES

- Bunda, M. A. An investigation of an extension of item sampling which yields individual scores. *Journal of Educational Measurement*, 1973, 10, 117-130.
- Feldt, L. S. and Forsyth, R. A. An examination of the context effect in item sampling. *Journal of Educational Measurement*, 1974, 11, 73-82.
- Sirotnik, K. A. Introduction to matrix sampling for the practitioner. In W. J. Popham (Ed.), *Evaluation in education: Current practices*. Berkeley, California: McCutchan Publishers, 1974.

DANIEL S. SHEEHAN

RONALD K. HAMBLETON

University of Massachusetts, Amherst

Ralph W. Tyler and Richard M. Wolf (Eds.) *Crucial Issues in Testing*. Berkeley, California: McCutchan, 1974. Pp. x + 170. \$9.25

Any book that presumes to summarize issues in a diverse professional field presents its authors with a choice between giving balanced treatment (equal time?) to competing viewpoints or offering a considered judgment based on a review of the arguments or evidence, putting them in a coherent framework leading to whatever evaluative conclusion has been reached. Tyler and Wolf have chosen the latter alternative. Even where articles by others have been used to reflect different views, a conclusion emerges.

Two questions need answering in the review of such a publication as this. First, did the authors choose the truly crucial issues? Second, how well did they summarize and generalize the state of the art?

The seven-part table of contents answers the first question affirmatively, covering (1) the testing of minority groups, (2) selective testing for higher education, (3) testing for grouping students for instruction, (4) criterion-referenced testing, (5) assessing the educational achievement of schools or school systems, "accountability," (6) testing to evaluate effectiveness of programs, methods, and materials of instruction, and (7) testing and the invasion of privacy. Inevitably some of these topics overlap or involve interaction, e.g., ability grouping and

the testing of minority groups, but the seven parts focus on definable areas of test use requiring attention by builders and users of tests in education without serious omission except possibly the use of tests in evaluating non-traditional acquisition of certifiable mastery of academic learning for high school or college credit.

Within the seven areas, then, the several sections may be evaluated seriatim. His proper concern regarding the testing situation's impact on minority individuals and their performance is clearly and temperately articulated by Robert Williams for the Association of Black Psychologists. It is treated comprehensively by Messick and Anderson, who admit the fairness of many of the criticisms, but conclude their rejection of the proposed moratorium on all "psychological" testing by pointing to the greater unfairness to be expected when stereotyped thinking in the dominant white group is not tempered by the objectivity of test performance. Robert Thorndike's sophisticated analysis of test fairness to groups should be read here if one has not encountered it previously. Use of a predictor variable to produce no greater group discrimination in selection than is found in criterion mastery after training is a powerful concept this reviewer finds a step forward from previous simpler concepts of fairness that ask only for comparable predictive validity.

The three excerpts from the 1970 report of the College Entrance Examination Board's Commission on Tests are commendably self-critical. One may wonder why something on service to public colleges, more characteristic of the American College Testing Program, was not included to show the problems of adaptation to the needs of institutions that have been drawn sooner and farther into open admissions. A useful historical function is performed by reviewing the trend within this reviewer's life-time in high school graduation and college attendance. Admission to "prestige" colleges is still an active concern of upwardly mobile middle-class families and influences the public mind. Edmund Gordon's brief for a more functional placement emphasis in the achievement tests is especially wholesome if some may find it too demanding of tests. One may point to considerable placement use of these tests, past and present, and expect supplementary information obtainable from and about applicants to be used to the full. The College Level Examination Program and the Advanced Placement Tests are obvious moves in a constructive direction.

Tyler's compact summary of the case against "ability grouping" is well put. A more comprehensive recent survey coauthored by this reviewer simply documents the views attributed to Heathers in greater detail. Homogeneity of achievement across even the strictly academic curriculum is chimerical, and grouping based on the assumption of such homogeneity may allow a few leaders to run ahead, but only at the expense of more substantial declines in the progress of stigmatized and understimulated (intellectually segregated) low track groups. Again, testing is not the chief culprit, but the *user* of tests who, if

deprived of tests, will make even worse classification schemes out of sheer prejudice and/or stereotyping.

Airasian and Madaus may be credited with a forceful, albeit somewhat polemical statement of the case for criterion-referenced testing. The strengths and merits of measuring qualitatively what behaviors a student can perform are clearly presented against a backdrop of the admittedly unimaginative use often made of even the best standardized tests of achievement. The villain is not the norm-referenced test per se, but rigidities of school operation that persist in hurrying all along at a uniform pace. Norm-referenced tests were welcomed and can be used as means of describing what traditional school arrangements engender. Criterion-referenced measurement promotes a better focus on the teaching-learning process than group measures based on the limited item samples of standardized achievement tests. But also there remain subareas of achievement not prescribable for mastery learning by all. The active, well-honed, highly-motivated mind will attain understandings beyond any specifiable set of outcomes in essentially domain-referenced areas, to use Gronlund's term. Criterion-referenced testing is a welcome addition to our armamentarium of measurement that will increasingly serve to define meaningful, attainable goals to the many we teach. It also has the virtue of goal-oriented evaluation, freed of unnecessary comparisons and contrasts. But let there continue to be recognized desirable levels of understanding, insight and even mastery not all will achieve, in addition to the minimum essentials specified in operational terms measurable by mastery tests.

Assessment approaches at national, state and local levels are discussed in turn. Highly efficient summaries of purpose, design and use are presented for the National Assessment of Educational Progress (NAEP). This type of assessment, which this reviewer remembers then Commissioner George Stoddard recommending to assembled schoolmen in New York State as long ago as 1943, is already in its first decade yielding the benefits of triennial cycles of assessment of ten areas at four age levels with other well-conceived breakdowns, geographical and otherwise.

State assessment, inevitably more diverse, is consequently more difficult to summarize. Unfortunately for the purposes of this publication there is some truth in Dyer and Rosenthal's statement that "state assessment programs are currently in a highly fluid state" so that "the facts and surmises presented in this (survey) report may well be out-of-date within a matter of months." Particularly serious are the conflicting purposes of testing programs in the same state, where one is long or short standing for providing individually reliable information to guide students and their teachers or counselors, while the other is introduced to assess the progress of the state system or definable subgroups over time. Such differences are being worked out, but produce frustrations and animosities meanwhile that take some time to

dissipate. An added note is that a recent Alabama study has provided a regression model for interpreting the achievement level of school systems relative to expectations based on economic indices, thereby allowing good teaching and learning in disadvantaged situations to show to advantage relative to expectations rather than unfavorably relative to national norms.

Wolf's call for local programs responsive to immediate demands for accountability includes an incidental account of the history of achievement testing since Rice's 1897 spelling studies that does justice to the major trends and counter trends. Perhaps the 1972 account this reviewer remembers of a Dallas (Texas) criterion-reverenced testing program built and normed within the system augurs a reconciliation of emphases now treated as antithetical.

Who better than Tyler could review the history and outline the current situation regarding use of tests in curriculum evaluation? The call is for specificity in measurement directed to the essential specificity of learning, a far cry from the bold measures of general qualities like excessive caution and over-generalization in "interpretation of data" tests. Students do develop abilities over time out of more specific directed learning. Generalizations are judicious syntheses of well-learned specifics; can they not do more than coexist peacefully, but build interactively on each other?

Wolf's discussion of the invasion of privacy via testing has a balance that suggests the possibility of clearly established ground rules to protect privacy without hobbling use of testing for educational diagnosis and research. It will take doing, but can be done.

In sum, *Crucial Issues in Testing* raises most of the critical issues to the point of visibility, is sometimes one-sidedly polemical, but elsewhere points the way to future realization of important new benefits and/or reconciliation of conflicting considerations. You'd better read it.

WARREN G. FINDLEY  
*University of Alabama*  
*Birmingham*



EDUCATIONAL and  
PSYCHOLOGICAL



MEASUREMENT

W. SCOTT GEHMAN, *Editor*

GERALDINE R. THOMAS, *Managing Editor*

WILLIAM B. MICHAEL, *Editor, Validity Studies and Computer Programs*

JOAN J. MICHAEL, *Assistant Editor, Validity Studies and Computer Programs*

MAX D. ENGELHART, *Book Review Editor*

LEWIS R. AIKEN, JR., *Assistant Book Review Editor*

FREDERIC KUDER, *Editor Emeritus*

#### BOARD OF COOPERATING EDITORS

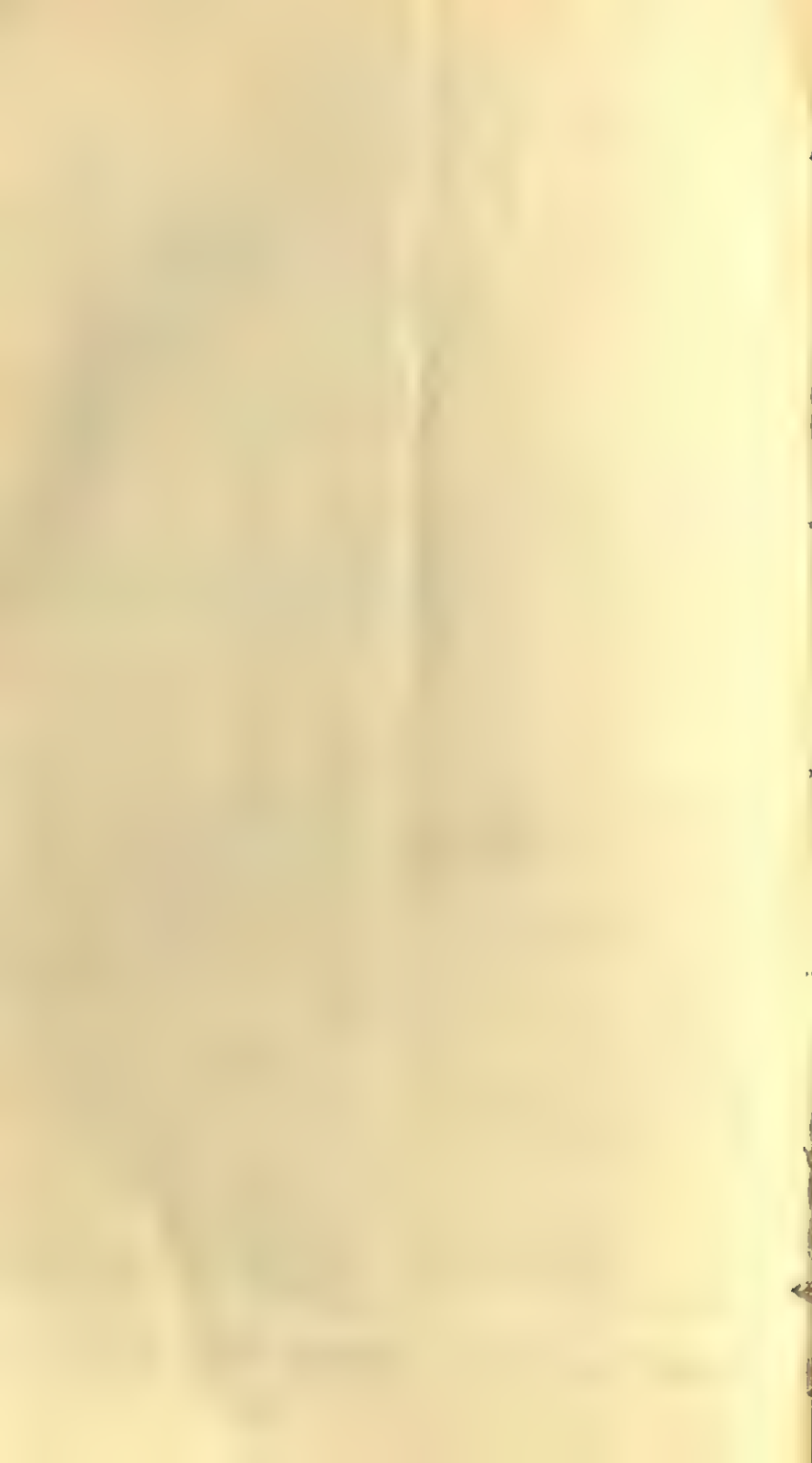
ROTHY C. ADKINS, *University of Hawaii*  
WIS R. AIKEN, JR., *University of Illinois*  
ROLD P. BECHTOLDT, *The University of Iowa*  
LLIAM V. CLEMANS, *American Institutes for Research*  
LIS D. COHEN, *University of Florida*  
THONY J. CONGER, *Duke University*  
ILS A. DAVIS, *Research Triangle Institute*  
ROLD A. EDGERTON, *Performance Research, Inc.*  
NE V GLASS, *University of Colorado*  
P. GUILFORD, *University of Southern California, Los Angeles*  
RK A. HORNADAY, *Babson College*  
HN E. HORROCKS, *The Ohio State University*  
RIL J. HOYT, *University of Minnesota*  
TON D. JACOBSON, *University of Virginia*  
SEPH C. JOHNSON, II, *Jackson State University*  
LIAM G. KATZENMEYER, *Duke University*  
BERT E. LANA, *Temple University*  
EDERIC M. LORD, *Educational Testing Service*  
DIE LUBIN, *Navy Medical Neuropsychiatric Research Unit, San Diego*

LOUIS L. MCQUITTY, *University of Miami, Coral Gables*  
HOWARD G. MILLER, *North Carolina State University at Raleigh*  
ROBERT L. MORGAN, *North Carolina State University at Raleigh*  
HENRY MOUGHAMIAN, *City Colleges of Chicago*  
DAVID NOVAK, *The Neuse Clinic, New Bern, N. C.*  
ELLIS B. PAGE, *University of Connecticut*  
NAMBURY S. RAJU, *Science Research Associates, Inc.*  
BEN H. ROMINE, JR., *University of North Carolina at Charlotte*  
THELMA G. THURSTONE, *University of Montana*  
WILLARD G. WARRINGTON, *Michigan State University*  
JOHN L. WASIK, *North Carolina State University at Raleigh*  
KINNARD WHITE, *University of North Carolina at Chapel Hill*  
JOHN E. WILLIAMS, *Wake Forest University*  
E. G. WILLIAMSON, *University of Minnesota*

VOLUME THIRTY-FIVE, NUMBER FOUR, WINTER 1975

820.60





# AUTHOR INDEX FOR VOLUME 35

<i>Page No.</i>	<i>Page No.</i>
Abbott, Robert D.....	371
Abrami, Philip C.....	885
Abu-Sayf, F. K.....	451
Adank, Richard.....	499
Adank, Richard.....	503
Aiken, Lewis R.....	181
Airasian, Peter W.....	809
Ansari, Z. A.....	1033
Archambault, Francis X.....	689
Bailit, Howard.....	843
Baird, Leonard L.....	941
Balcar, Karel R.....	175
Baldauf, Richard B., Jr.....	723
Barker, William B.....	427
Batlis, Nick.....	447
Baum, Dale D.....	487
Bean, Andrew.....	963
Belcastro, Frank P.....	957
Best, Deborah L.....	3
Bland, Patricia C.....	175
Boswell, Donna A.....	3
Boyle, David.....	897
Braver, Sanford L.....	283
Breen, Lawrence J.....	885
Brennan, Robert L.....	779
Brunza, J. Jay.....	73
Bruvold, William H.....	605
Burns, John A.....	869
Butcher, James N.....	393
Butterworth, Thomas W.....	987
Carroll, Robert M.....	541
Carter-Saltzman, Louise.....	427
Chamberlain, Howard.....	721
Chansky, Norman M.....	947
Chardos, Steve.....	353
Chissom, Brad S.....	461
Christiaans, H.H.C.M.....	969
Cliff, Norman.....	671
Cliff, Norman.....	675
Coletta, Anthony J.....	415
Conger, Anthony J.....	847
Conger, Judith Cohen.....	847
Cook, Daniel W.....	529
Cooper, Martin.....	303
Covert, Robert.....	495
Covert, Robert W.....	947
Crehan, Kevin D.....	97
Croskey, Frank L.....	735
Cureton, E.....	47
D'Agostino, Ralph B.....	47
Dawis, René V.....	51
Dawis, René V.....	325
Devine, Bernard.....	797
Dombrower, Jule.....	993
Doughty, Gavin, Jr.....	733
Dubois, Bernard.....	869
Dziuban, Charles D.....	539
Echternacht, Gary.....	307
Eliot, John.....	975
Favero, Jane.....	993
Feild, Hubert S.....	171
Foster, Glen G.....	1023
Fugita, Stephen S.....	745
Gable, Robert K.....	415
Games, Paul A.....	147
Gheorghiu, V. A.....	341
Goolsby, Thomas M., Jr.....	507
Graves, Deborah J.....	3
Greene, John F.....	689
Gross, Alan L.....	143
Halperin, Silas.....	159
Hamilton, J. Ogden.....	915
Harper, Frank B. W.....	905
Hayes, Marjorie.....	495
Haynes, Jack R.....	107
Heeler, Roger M.....	255
Helwig, Loren D.....	507
Hendel, Darwin D.....	865
Hillery, Joseph M.....	745
Hodapp, V.....	341
Hofmann, Richard J.....	191
Hofmann, Richard J.....	621
Hoogstraten, Joh.....	969
Huck, Schuyler W.....	789
Hunfer, Sara.....	393
Jackson, Douglas N.....	361
Jackson, Douglas N.....	663
James, Mark.....	185
Jaspen, Nathan.....	697
Jaspen, Nathan.....	701
Jensen, Joan M.....	1011
Johnson, Richard W.....	951
Jones, Michael P.....	729
Jones, Phillip D.....	821
Kaiser, Henry F.....	31
Karabinus, R. A.....	277
Karweit, Nancy.....	153
Katzenmeyer, William G.....	19
Kaufman, Alan S.....	641
Kaufman, Gary G.....	821
Keating, Daniel P.....	657
Kehoe, Jerard F.....	675
Khan, S. B.....	835
Kirk, Kenneth W.....	951
Koch, Valerie L.....	239
Koch, Valerie L.....	751

<i>Page No.</i>	<i>Page No.</i>		
Koehler, Roger A.....	97	Ohvall, Richard A.....	951
Koslowsky, Meni.....	843	Olejnik, Stephen.....	37
Krauth, J.....	231	Oles, Henry J.....	437
Krus, David J.....	175	Overall, John E.....	393
Kunce, Joseph T.....	529	Pandey, Tej N.....	567
Lanier, Doris.....	461	Pascale, Pietro J.....	733
Lawlis, G. Frank.....	313	Pearce, W. Barnett.....	115
Leventhal, Les.....	885	Pedersen, L. C.....	509
Levy, Kenneth J.....	599	Pedrini, D. T.....	717
Levy, Kenneth J.....	793	Peeples, Thomas O.....	539
Lewis, John.....	465	Perney, Jan.....	983
Lewis, John.....	467	Perry, Raymond P.....	885
Lewis, John.....	499	Phelps, Fred D.....	455
Lewis, John.....	503	Pomerantz, Michael.....	379
Lieblich, Amia.....	473	Porter, Andrew C.....	37
Lienert, G. A.....	231	Prevatt, Truman W.....	153
Lissitz, Robert W.....	353	Price, Karl F.....	911
Lovell, Constance.....	477	Price, Lewis.....	975
Ludwig, C. M.....	341	Pugh, Richard C.....	73
Lundberg, Ulf.....	797	Racioppo, Vincent.....	733
Macready, George B.....	583	Rafacz, Bernard A.....	167
Madaus, George F.....	809	Ramraz, Rachel.....	683
Majors, Gene W.....	1005	Ray, Michael L.....	255
Many, Margaret A.....	1017	Redburn, F. Stevens.....	767
Many, Wesley A.....	1017	Reilly, Richard R.....	613
Mårdberg, Bertil.....	163	Reynolds, Thomas J.....	671
Mason, Emanuel.....	495	Richards, James M., Jr.....	153
Mattson, Linda A.....	3	Richards, Leo.....	477
McCook, William M.....	689	Richards, Leo.....	993
McCoy, Rose E.....	935	Riley, F. Terrill.....	921
McDonald, John F.....	929	Roberts, Dennis M.....	921
McLaughlin, Donald H.....	79	Rucker, Margaret H.....	319
McNary, Susan.....	477	Sabatino, David A.....	1023
McPherson, Michael S.....	929	Scarr-Salapatek, Sandra.....	427
McQuitty, Louis L.....	239	Schmidt, Carl R.....	1023
McQuitty, Louis L.....	751	Schoenfeldt, Lyle F.....	171
Michael, Joan J.....	405	Schriesheim, Chester A.....	189
Michael, Joan J.....	1005	Schroeder, Lee L.....	685
Michael, Joan J.....	1011	Schultz, Charles B.....	379
Michael, William B.....	31	Schutz, Howard G.....	319
Michael, William B.....	185	Seay, Thomas A.....	921
Michael, William B.....	405	Shinar, Maya.....	473
Michael, William B.....	477	Shine, Lester C., II.....	535
Michael, William B.....	987	Shirkey, Edwin C.....	539
Michael, William B.....	993	Shoemaker, David M.....	567
Michael, William B.....	1011	Shor, Eli.....	683
Miller, Douglas E.....	529	Shurling, James.....	979
Montanelli, Richard G., Jr.....	195	Simono, R. B.....	401
Morgan, Ronald R.....	387	Simpson, Kenneth C.....	897
Morris, John D.....	707	Skinner, Harvey A.....	663
Morrison, Thomas L.....	119	Slakter, Malcolm J.....	97
Mueller, Daniel J.....	135	Smetana, Frederick O.....	679
Nagi, John L.....	471	Smith, Robert A.....	185
Nevo, Barukh.....	683	Smith, Robert A.....	405
Nordholm, Lena A.....	541	Sockloff, Alan I.....	267

	<i>Page No.</i>		<i>Page No.</i>
Stenner, A. Jackson .....	19	Waters, Carrie Wherry .....	447
Stofflet, Frederick .....	1029	Waters, L. K. ....	447
Straton, Ralph G. ....	555	Weiner, Max .....	455
Sutton, Cary O. ....	789	Welch, Margaret .....	467
Sweney, Arthur B. ....	313	Wen, Shih-sung .....	935
Thornton, Billy W. ....	735	Wherry, Robert J., Sr. ....	189
Tinsley, Howard E. A. ....	325	Whitley, Susan E. ....	51
Tittle, Carol Kehr .....	455	Wiebe, Bernie .....	115
Tokar, Edward B. ....	1029	Williams, John E. ....	3
Tritchler, D. L. ....	717	Wimberley, Ronald C. ....	693
Turner, Charles F. ....	667	Woodbury, Roger .....	979
Van Fleet, David D. ....	721	Woods, Elinor M. ....	809
Vegelius, Jan. ....	711	Wotruba, Thomas R. ....	911
Vegelius, Jan. ....	713	Yoshida, Roland K. ....	729
Vitaliano, Peter P. ....	159	Za'rour, George I. ....	451
Ward, William C. ....	87		

U. S. POSTAL SERVICE  
**STATEMENT OF OWNERSHIP, MANAGEMENT AND CIRCULATION**  
(Act of August 12, 1970: Section 3685 Title 39 United States Code)

1. TITLE OF PUBLICATION <b>Educational and Psychological Measurement</b>	2. DATE OF FILING <b>October 1, 1975</b>
3. FREQUENCY OF ISSUE <b>Quarterly</b>	3A. ANNUAL SUBSCRIPTION PRICE <b>\$20.00</b>

4. LOCATION OF KNOWN OFFICE OF PUBLICATION (Street, city, county, state and ZIP code) (Not printers)  
**3121 Cheek Road, Durham, N. C. 27704**

5. LOCATION OF THE HEADQUARTERS OR GENERAL BUSINESS OFFICES OF THE PUBLISHERS (Not printers)  
**3121 Cheek Road, Durham, N. C. 27704**

6. NAMES AND ADDRESSES OF PUBLISHER, EDITOR, AND MANAGING EDITOR

PUBLISHER (Name and address)

**G. Frederic Kuder, Box 6907 College Station, Durham, N. C. 27708**

EDITOR (Name and address)

**W. Scott Gehman, Box 6907 College Station, Durham, N. C. 27708**

MANAGING EDITOR (Name and address)

**Geraldine R. Thomas, 3121 Cheek Road, Durham, N. C. 27704**

7. OWNER (If owned by a corporation, its name and address must be stated and also immediately thereunder the names and addresses of stockholders owning or holding 1 percent or more of total amount of stock. If not owned by a corporation, the names and addresses of the individual owners must be given. If owned by a partnership or other unincorporated firm, its name and address, as well as that of each individual must be given.)

NAME

ADDRESS

**G. Frederic Kuder (Owner)**

**Box 6907 College Station, Durham, N. C. 27708**

8. KNOWN BONDHOLDERS, MORTGAGEES, AND OTHER SECURITY HOLDERS OWNING OR HOLDING 1 PERCENT OR MORE OF TOTAL AMOUNT OF BONDS, MORTGAGES OR OTHER SECURITIES (If there are none, so state)

NAME

ADDRESS

**None**

9. FOR OPTIONAL COMPLETION BY PUBLISHERS MAILING AT THE REGULAR RATES (Section 132.121, Postal Service Manual, 39 U. S. C. 3626 provides in pertinent part: "No person who would have been entitled to mail matter under former section 4359 of this title shall mail such matter at the rates provided under this subsection unless he files annually with the Postal Service a written request for permission to mail matter at such rates.")

In accordance with the provisions of this statute I hereby request permission to mail the publication named in Item 1 at the reduced postage rates presently authorized by 39 U. S. C. 3626

Signature and title of editor, publisher, business manager, or owner

*Geraldine R. Thomas, Managing Editor*

10. FOR COMPLETION BY NONPROFIT ORGANIZATIONS AUTHORIZED TO MAIL AT SPECIAL RATES (Section 132.122 Postal Service Manual) (Check one)

The purpose, function, and nonprofit status of this organization and the exempt status for Federal income tax purposes

☐ Have not changed during preceding 12 months

☐ Have changed during preceding 12 months

(If changed, publisher must submit explanation of change with this statement.)

11. EXTENT AND NATURE OF CIRCULATION

AVERAGE NO. COPIES EACH ISSUE DURING PRECEDING 12 MONTHS

ACTUAL NUMBER OF COPIES OF SINGLE ISSUE PUBLISHED NEAREST TO FILING DATE

A. TOTAL NO. COPIES PRINTED (Net Press Run)

B. PAID CIRCULATION

1. SALES THROUGH DEALERS AND CARRIERS, STREET VENDORS AND COUNTER SALES

2. MAIL SUBSCRIPTIONS

C. TOTAL PAID CIRCULATION

D. FREE DISTRIBUTION BY MAIL, CARRIER OR OTHER MEANS SAMPLES COMPLIMENTARY, AND OTHER FREE COPIES

E. TOTAL DISTRIBUTION (Sum of C and D)

F. COPIES NOT DISTRIBUTED

1. OFFICE USE, LEFT OVER, UNACCOUNTED, SPOILED AFTER PRINTING

2. RETURNS FROM NEWS AGENTS

G. TOTAL (Sum of B & F—should equal net press run shown in A)

3628

3710

I certify that the statements made by me above are correct and complete

SIGNATURE OF EDITOR, PUBLISHER, BUSINESS MANAGER, OR OWNER

*Geraldine R. Thomas, Managing Editor*





